PAC-Bayesian Bounds for Learning LTI-ss systems with Input from Empirical Loss

Deividas Eringis [1] **John Leth** [1] **Zheng-Hua Tan** [1] **Rafal Wisniewski** [1] **Mihaly Petreczky** [2]

Abstract

In this paper we derive a Probably Approximately Correct(PAC)-Bayesian error bound for linear time-invariant (LTI) stochastic dynamical systems with inputs. Such bounds are widespread in machine learning, and they are useful for characterizing the predictive power of models learned from finitely many data points. In particular, the bound derived in this paper relates future average prediction errors with the prediction error generated by the model on the data used for learning. In turn, this allows us to provide finite-sample error bounds for a wide class of learning/system identification algorithms. Furthermore, as LTI systems are a sub-class of recurrent neural networks (RNNs), these error bounds could be a first step towards PAC-Bayesian bounds for RNNs.

1. Introduction

Motivation PAC and PAC-Bayesian bounds have been a major tool for analyzing learning algorithms. They provide bounds on the generalization error in terms of the empirical error, in a manner which is independent of the learning algorithm. Hence, these bounds can be used to analyze and explain a wide variety of learning algorithms. In particular, PAC-Bayesian error bounds turned out to be useful for providing non-vacuous error bounds for neural networks (Dziugaite & Roy, 2017).

While there is a wealth of literature on PAC (Shalev-Shwartz & Ben-David, 2014) and PAC-Bayesian (Alquier, 2021; Guedj, 2019), bounds for static models, much less is known on dynamical systems.

*Equal contribution [1]Dept. of Electronic Systems, Aalborg University, Denmark [2]Laboratoire Signal et Automatique de Lille (CRIStAL), Lille, France. Correspondence to: Deividas Eringis <der@es.aau.dk>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

Contribution In this paper we consider stochastic LTI state-space representations (LTI systems for short) in innovation form. In accordance with the standard practice in system identification, we view stochastic LTI systems as predictors, which take past inputs and outputs and generate predictions for the current output. We assume that the data used for learning are generated by stochastic LTI systems in innovation form. Learning/identifying an LTI system then amounts to finding the predictor, which results in the smallest prediction error for the training data, i.e., the smallest *empirical loss*. However, for decision making (fault detection,control, etc.), the quality of the learned model is determined by the *generalization error*, i.e., the average prediction error for future, unseen data. The PAC-Bayesian bound of this paper says that with a high probability (probability at least $1-\delta$), the generalization error is smaller than the empirical loss plus an error term. The error term depends on the number of data points N and on parameter (learning rate λ). In this paper we provide explicit formulas for the error term. We show that the error term converges to a constant as $N \to \infty$. The constant depends on the confidence level δ and the distance between prior and posterior densities on models. If we assume that the data used for learning is generated by an LTI system with *bounded noise*, we can show that the error term converges to 0 as $N \to \infty$. The rate of convergence is $O(\frac{1}{\sqrt{N}})$, which is consistent with most of finite-sample bounds available in the literature for various, not necessarily LTI, models. This suggests that the obtained error bound is likely to be asymptotically sharp for bounded signals.

Related work PAC bounds for sub-classes of linear dynamical systems in autoregressive form and with bounded signals were proposed in (Campi & Weyer, 2002; Vidyasagar & Karandikar, 2006). In (Alquier & Wintenberger, 2012; Alquier et al., 2013) PAC-Bayesian bounds for autoregressive models without exogenous inputs were considered, and the variables were either assumed to be bounded or the loss function was assumed to be Lipschitz.

In contrast to (Campi & Weyer, 2002; Vidyasagar & Karandikar, 2006; Alquier et al., 2013; Alquier & Wintenberger, 2012), we consider state-space models with inputs, we also handle unbounded variables (although the results are weaker than for the bounded case) and quadratic loss func-

tions. However, results from (Alquier et al., 2013; Alquier & Wintenberger, 2012) were used in this paper. Another related work is (Massucci et al., 2021), where PAC bounds for switched autoregressive systems were derived, but again those results do not apply to stochastic LTI state-space representations.

Recently, several publications on finite-sample bounds for learning linear dynamical systems were derived, without claiming completeness (Simchowitz et al., 2019; Simchowitz, 2021; Oymak & Ozay, 2022; Lale et al., 2020; Foster & Simchowitz, 2020; Hazan et al., 2018; Tsiamis & Pappas, 2019; Sarkar et al., 2021). First, all the cited papers propose a bound which is valid only for models generated by a specific learning algorithm. In particular, these bounds do not relate the generalization loss with the empirical loss for arbitrary models, i.e., they are not PAC(-Bayesian) bounds. This means that in contrast to the results of this paper, the bounds of the cited papers cannot be used for analyzing algorithms others than for which they were derived. Second, many of the cited papers do not derive bounds on the infinite horizon prediction error. PAC bounds for recurrent neural networks, of which LTI state-space representations are a subclass, were developed in (Koiran & Sontag, 1998; Sontag, 1998; Chen et al., 2020) using VC dimension, and in (Joukovsky et al., 2021; Chen et al., 2020) using Rademacher complexity, and in (Zhang, 2006; Dziugaite & Roy, 2017) using PAC-Bayesian bounds approach. However, all the cited papers assume noiseless models, a fixed number of time-steps, that the training data are i.i.d sampled time-series, and the signals are bounded. In contrast, we consider (1) noisy models, (2) prediction error defined on infinite time horizon, (3) only one single time series available for training data, and (4) we allow unbounded signals.

In (Eringis et al., 2021) PAC-Bayesian error bounds were developed for autonomous LTI state-space systems without exogenous input. In contrast to (Eringis et al., 2021), in the current paper we consider systems with exogenous inputs. Moreover, the error bound of this paper is much tighter than that of (Eringis et al., 2021): in contrast to (Eringis et al., 2021), with the growth of the number of observations, the error bounds of this paper converge either to zero (in the case of bounded innovation noise) or to a constant involving KL-divergence. Finally, the proof technique is completely different from that of (Eringis et al., 2021).

Paper Outline We start by defining the problem formulation in Section 2, where all the assumptions and important quantities are defined. Then we will discuss the PAC-Bayesian framework in Section 3, then we will present the main results of the paper in Section 4, then we will present some auxiliary results for systems driven by bounded noise in Section 5, Finally, a short numerical example is presented in Section 6.

2. Problem formulation

Notation and terminology

We occasionally use $\triangleq$ to denote "defined by". Let $\mathbf{F}$ denote a σ-algebra on the set Ω and $\mathbf{P}$ be a probability measure on $\mathbf{F}$. Unless otherwise stated all probabilistic considerations will be with respect to the probability space $(\Omega, \mathbf{F}, \mathbf{P})$, and we let $\mathbf{E}(\mathbf{z})$ denote expectation of the stochastic variable $\mathbf{z}$. We use bold face letters to indicate stochastic variables/processes. Each euclidean space is associated with the topology generated by the 2-norm $\|\cdot\|_2$, and the Borel σ-algebra generated by the open sets. The induced matrix 2-norm is also denoted $\|\cdot\|_2$. We say that a random variable $\mathbf{z}$ taking values in $\mathbb{R}^n$ is essentially bounded, if for some constant $C > 0$, $\|\mathbf{z}\|_2 < C$ holds with probability one.

A *stochastic linear-time invariant (LTI) systems with inputs in state-space form* (Lindquist & Picci, 2015, Chapter 17) is a dynamical system of the form

$$\begin{aligned} \mathbf{x}(t+1) &= A\mathbf{x}(t) + B\mathbf{u}(t) + \boldsymbol{\nu}(t), \\ \mathbf{y}(t) &= C\mathbf{x}(t) + D\mathbf{u}(t) + \boldsymbol{\eta}(t) \end{aligned} \tag{1}$$

defined for all $t \in \mathbb{Z}$, where A, B, C, D are $n \times n$, $n \times n_{\mathbf{u}}$, $n_{\mathbf{y}} \times n$ and $n_{\mathbf{y}} \times n_{\mathbf{u}}$ matrices respectively, A is a Schur matrix (a square matrix with all its eigenvalues inside the unit disk), $\boldsymbol{\nu}, \boldsymbol{\eta}$ are zero-mean Gaussian i.i.d processes, $\mathbf{u}$, $\mathbf{x}$, are zero-mean stationary Gaussian processes, $\mathbf{u}(t)$ and $\left[\boldsymbol{\eta}^T(t), \boldsymbol{\nu}^T(t)\right]^T$ are independent, and $\mathbf{x}(t)$ and $\left[\boldsymbol{\nu}^T(t), \boldsymbol{\eta}^T(t)\right]^T$ are independent. The process $\mathbf{x}$ is called the state process, $\boldsymbol{\nu}$ is called the process noise and $\boldsymbol{\eta}$ is the measurement noise. If B, D are absent from (1), then we say that (1) is an *autonomous stochastic LTI system*

Let us fix stochastic processes $\mathbf{y}(t) \in \mathbb{R}^{n_{\mathbf{y}}}$, and $\mathbf{u}(t) \in \mathbb{R}^{n_{\mathbf{u}}}$, that share a time axis $t \in \mathbb{Z}$, that is, for any $t \in \mathbb{Z}$, $\mathbf{y}(t) : \Omega \to \mathbb{R}^{n_{\mathbf{y}}}; \omega \mapsto \mathbf{y}(t)(\omega)$, and $\mathbf{u}(t) : \Omega \to \mathbb{R}^{n_{\mathbf{u}}}; \omega \mapsto \mathbf{u}(t)(\omega)$ are random vectors on $(\Omega, \mathbf{F}, \mathbf{P})$. The goal is to estimate $\mathbf{y}(t)$ from current and past values of $\mathbf{u}(t)$, for this we need a structure connecting $\mathbf{y}(t)$ and $\mathbf{u}(t)$, thus we have

Assumption 2.1. Let $\mathbf{y}(t)$ and $\mathbf{u}(t)$ be generated by an autonomous stochastic LTI system

$$\mathbf{x}(t+1) = A_g\mathbf{x}(t) + K_g\mathbf{e}_g(t), \tag{2a}$$

$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix} = C_g\mathbf{x}(t) + \mathbf{e}_g(t) \tag{2b}$$

where $A_g \in \mathbb{R}^{n\times n}, K_g \in \mathbb{R}^{n\times m}, C_g \in \mathbb{R}^{m\times n}$ for $n > 0$, $m = n_{\mathbf{y}} + n_{\mathbf{u}} \geq 2$. Furthermore, we require that

- A_g and $A_g - K_gC_g$ are Schur (all its eigenvalues are inside the open unit circle), and

- $\mathbf{e}_g$ is an i.i.d process and $\mathbf{e}_g(t)$ is a zero mean sub-Gaussian variable for all $t \in \mathbb{Z}$ with covariance $\mathbf{E}[\mathbf{e}_g(t)\mathbf{e}_g^T(t)] = Q_e$,
- $\mathbf{x}(t)$ has finite variance, and $\mathbf{e}_g(t)$ is independent of $\mathbf{x}(t)$, for all $t \in \mathbb{Z}$.

We identify the system (2) with the tuple $\Sigma_{gen} \triangleq (A_g, K_g, C_g, I)$

Note: For learning, we assume to have the training data set $\mathcal{D}_N = \{\{\mathbf{y}(s)(\omega), \mathbf{u}(s)(\omega)\}\}_{s=0}^{N-1}$, i.e. a single trajectory of $[\mathbf{y}^T(t), \mathbf{u}^T(t)]^T$, but no knowledge of the matrices A_g, K_g, C_g and noise process $\mathbf{e}_g$. The system (2) only defines the assumptions on the data generating process.

The goal is to use the past and present of $\mathbf{u}(t)$, or past of $\mathbf{y}(t)$, to estimate $\mathbf{y}(t)$. Note that $\mathbf{y}$ and $\mathbf{u}$ are stationary processes by (Caines, 1988, Theorem 1.4). Moreover, from classical theory of LTI systems it follows that $\mathbf{y}(t)$ and $\mathbf{u}(t)$, $t \in \mathbb{Z}$ are essentially bounded if the noise $\mathbf{e}_g(s)$ is essentially bounded for all $s \in \mathbb{Z}$

We wish to consider LTI predictors,

$$\hat{\mathbf{x}}(t+1) = \hat{A}\hat{\mathbf{x}}(t) + \hat{B}\mathbf{u}(t) + \hat{L}\mathbf{y}(t),\ \hat{\mathbf{x}}(0) = 0 \quad (3a)$$

$$\hat{\mathbf{y}}(t) = \hat{C}\hat{\mathbf{x}}(t) + \hat{D}\mathbf{u}(t) \quad (3b)$$

where matrices $\hat{A}, \hat{B}, \hat{L}, \hat{C}, \hat{D}$ are of appropriate size, and $\hat{A}$ is Schur (all its eigenvalues are inside the unit disk).

Note: In this paper, we will allow a more general form of predictors, where $\hat{L}$ can be set to 0, i.e. we may wish to estimate $\mathbf{y}(t)$ only from measurements $\mathbf{u}(t)$, when past values of the process $\mathbf{y}(t)$ is not available. In order to accommodate this let us define a stochastic process $\mathbf{w}(t) \in \mathbb{R}^{n_\mathbf{w}}$, by two cases

- $\mathbf{w}(t) = \begin{bmatrix}\mathbf{y}^T(t) & \mathbf{u}^T(t)\end{bmatrix}^T$, $n_\mathbf{w} = n_\mathbf{y} + n_\mathbf{u}$
- $\mathbf{w}(t) = \mathbf{u}(t)$, $n_\mathbf{w} = n_\mathbf{u}$

Note that, one can define $\mathbf{w}(t)$, to consist of some of the components of $\mathbf{y}(t)$, i.e. $\mathbf{w}(t)$ does not need to contain all of $\mathbf{y}$.

Class of predictors (hypotheses) In this paper, we will be interested in the following hypothesis class, consisting of predictors realizable by LTI systems.

Assumption 2.2 (Parameterised hypothesis class)**.** The hypothesis class $\mathcal{F}$ is a parametrized set of LTI predictors, with $\Sigma(\theta) = (\hat{A}(\theta), \hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta))$:

$$\hat{\mathbf{x}}(t+1) = \hat{A}(\theta)\hat{\mathbf{x}}(t) + \hat{B}(\theta)\mathbf{w}(t),\ \hat{\mathbf{x}}(0) = 0, \quad (4a)$$

$$f_{\Sigma(\theta)}(\{\mathbf{w}(s)\}_{s=0}^t) = \hat{C}(\theta)\hat{\mathbf{x}}(t) + \hat{D}(\theta)\mathbf{w}(t). \quad (4b)$$

$$\mathcal{F} = \{f_{\Sigma(\theta)} \mid \gamma(\hat{A}(\theta)) < 1,\ \theta \in \Theta\}$$

with $\gamma(\hat{A}(\theta))$ the spectral radius of $\hat{A}(\theta)$, i.e. the largest modulus of eigenvalues of $\hat{A}(\theta)$. Set $\Theta \subset \mathbb{R}^{n_\theta}$ is a compact set, and $\hat{A}(\theta)$,$\hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta)$ are continuous functions of θ taking values in the sets of $\hat{n} \times \hat{n}$, $\hat{n} \times n_\mathbf{w}$, $n_\mathbf{y} \times \hat{n}$ and $n_\mathbf{y} \times n_\mathbf{w}$ matrices respectively. If $\mathbf{w}(t) = [\mathbf{y}^T(t), \mathbf{u}^T(t)]^T$, then $\hat{D} = [0, \hat{D}_\mathbf{u}]$ for some $n_\mathbf{y} \times n_\mathbf{u}$ matrix $\hat{D}_\mathbf{u}$, i.e., $\hat{D}\mathbf{w}(t)$ depends only on $\mathbf{u}(t)$[1].

We will identify the system (4) with the tuple $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$. For the sake of notation, throughout the paper we will use f, to denote $f_{\Sigma(\theta)}$, for some arbitrary $\theta \in \Theta$.

Under assumption 2.2, we can use probability densities on the set of predictors $\mathcal{F}$. The latter will be essential for using the PAC-Bayesian framework.

Next, we define the notions of empirical and generalization loss for predictors which are realized by LTI systems.

Assumption 2.3 (Quadratic loss function)**.** We will consider *quadratic loss functions* $\ell : \mathbb{R}^{n_\mathbf{y}} \times \mathbb{R}^{n_\mathbf{y}} \ni (y, y') \mapsto \|y - y'\|_2^2 = (y - y')^T(y - y') \in [0, \infty)$.

The empirical loss of a predictor for the data $\mathcal{D}_N = \{\mathbf{y}(t), \mathbf{w}(t)\}_{t=0}^N$ is defined as follows: we define the random variable

$$\hat{\mathbf{y}}_f(t \mid s) \triangleq f(\mathbf{w}(s), \ldots, \mathbf{w}(t))$$

which represents the estimate of $\mathbf{y}(t)$ based on random variables $\{\mathbf{w}(s), \ldots, \mathbf{w}(t)\}$. The *empirical loss for a predictor* f and processes $(\mathbf{y}, \mathbf{w})$ is then defined by

$$\hat{\mathcal{L}}_N(f) \triangleq \frac{1}{N}\sum_{i=0}^{N-1} \ell(\hat{\mathbf{y}}_f(i \mid 0), \mathbf{y}(i)). \quad (5)$$

The definition of the generalization loss is a bit more involved. Namely, we are using varying number of inputs for predictions and hence the expectation $\mathbf{E}[\ell(\hat{\mathbf{y}}_f(t \mid 0), \mathbf{y}(t))]$ depends on t. This will hold true even if the processes $\mathbf{y}$ and $\mathbf{w}$ are stationary. Note that this issue is specific for state-space models: autoregressive models always use the same number of inputs to make a prediction. In this paper we will opt for looking at the case when the size of the past used for the prediction is infinite.

Lemma 2.4 ((Hannan & Deistler, 1988))**.** *The limit* $\hat{\mathbf{y}}_f(t) = \lim_{s\to-\infty} \hat{\mathbf{y}}_f(t \mid s)$ *exists in the mean-square sense for all* t*, the process* $\hat{\mathbf{y}}_f(t)$ *is stationary, and* $\mathbf{E}[\ell(\hat{\mathbf{y}}_f(t), \mathbf{y}(t))] = \lim_{s\to-\infty} \mathbf{E}[\ell(\hat{\mathbf{y}}_f(t \mid s), \mathbf{y}(t))]$.

This motivates us to introduce the quantity

$$\mathcal{L}(f) = \mathbf{E}[\ell(\hat{\mathbf{y}}_f(t), \mathbf{y}(t))] = \lim_{s\to-\infty} \mathbf{E}[\ell(\hat{\mathbf{y}}_f(t \mid s), \mathbf{y}(t))]$$

[1]The latter assumption is necessary, since otherwise we would be using the components of $\mathbf{y}(t)$ to predict $\mathbf{y}(t)$, which is not meaningful.

which is called the *generalization loss* of the predictor f when applied to process $(\mathbf{y}, \mathbf{w})$.

Intuitively, $\hat{\mathbf{y}}_f(t)$ can be interpreted as the prediction of $\mathbf{y}(t)$ generated by the predictor f based on all (infinite) past and present values of $\mathbf{w}$. As stated in Lemma 2.4 we consider the special case when $\hat{\mathbf{y}}_f(t)$ is the mean-square limit of $\hat{\mathbf{y}}_f(t \mid s)$ as $s \to -\infty$. Clearly, for large enough $t - s$, the empirical loss, is close to the generalization loss. In fact, it is standard practice in learning dynamical systems (Ljung, 1999) to use $\mathcal{L}(f)$ as the measure of fitness of the predictor. With these definitions in mind, the learning problem considered in this paper can be stated as follows.

Problem 2.1 (Learning problem). Compute a predictor $f \in \mathcal{F}$ from a sample $\mathcal{D}_N = \{\mathbf{y}(t)(\omega), \mathbf{w}(t)(\omega)\}_{t=0}^{N}$ of the random variables $\{\mathbf{y}(t), \mathbf{w}(t)\}_{t=0}^{N}$ such that the generalization loss $\mathcal{L}(f)$ is small.

Remark 2.5 (Relationship with parameter estimation). The learning problem above can be interpreted as a parameter estimation problem as follows. Assume that there is no feedback from $\mathbf{y}$ to $\mathbf{u}$ in the sense of (Lindquist & Picci, 2015, Definition 17.1.1). Then from (Eringis et al., 2022) it follows that the data generator Σ_{gen} gives rise to a predictor $f(\Sigma_{gen})$ with $\mathbf{w} = \begin{bmatrix}\mathbf{y}^T & \mathbf{u}^T\end{bmatrix}^T$, such that generalization error $\mathcal{L}(f(\Sigma_{gen}))$ is the *smallest* possible among all the predictors. Moreover, the correspondence between the matrices of Σ_{gen} and $f(\Sigma_{gen})$ is one-to-one and continuous. Hence, if the predictor $f(\Sigma_{gen})$ arising from the data generator Σ_{gen} belongs to the hypothesis class, then the solution of the learning problem will be the predictor $f(\Sigma_{gen})$ which arises from the data generator, moreover, the matrices of the predictor can be used to compute the matrices of the data generator (Eringis et al., 2022). That is, the problem of finding a predictor with minimal generalization error is equivalent to finding the data generator. Moreover, if the hypothesis class satisfies some regularity conditions (e.g., identifiability, etc.), then any predictor with a sufficiently small generalization error will have matrices which are close to the matrices of the optimal predictor, and hence can be used to compute an approximation of the data generator.

3. PAC-Bayesian Framework

Below we recall the PAC-Bayesian framework. To this end, let B_Θ be the σ-algebra of Lebesque-measurable subsets of the parameter set $\Theta \subseteq \mathbb{R}^{n_\theta}$, and m denote the Lebesque measure on $\mathbb{R}^{n_\theta}$. We then define

$$\underset{f\sim\rho}{E} g(f) \triangleq \int_{\theta\in\Theta} \rho(\theta) g(f_{\Sigma(\theta)}) dm(\theta) \tag{6}$$

with ρ a probability density function on the measure space (Θ, B_θ, m), and $g : \mathcal{F} \to \mathbb{R}$ a map such that $\Theta \ni \theta \mapsto g(f_{\Sigma(\theta)})$ is measurable and absolutely integrable. The essence of the PAC-Bayesian approach is to prove that for any density π on $\mathcal{F}$, and any $\delta \in (0, 1]$,

$$P\left(\left\{\omega \in \Omega \mid \forall \hat{\rho} \in \mathcal{M}_\pi : \underset{f\sim\hat{\rho}}{E}\mathcal{L}(f) \leq \underset{f\sim\hat{\rho}}{E}\hat{\mathcal{L}}_N(f)(\omega) + r_N(\lambda, \hat{\rho}, \delta)\right\}\right) > 1 - \delta, \tag{7}$$

i.e., the inequality $\underset{f\sim\hat{\rho}}{E}\mathcal{L}(f) \leq \underset{f\sim\hat{\rho}}{E}\hat{\mathcal{L}}_N(f) + r_N(\lambda, \hat{\rho}, \delta)$ holds with probability at least $1 - \delta$, where

$$r_N(\lambda, \hat{\rho}, \delta) = \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \Psi_\pi(\lambda, N)\right], \tag{8}$$

where $\lambda > 0$ and $D_{\mathrm{KL}}(\hat{\rho} \mid \pi) \triangleq E_{f\sim\hat{\rho}} \ln\frac{\hat{\rho}(f)}{\pi(f)}$ is the KL-divergence between π and $\hat{\rho}$, and

$$\Psi_\pi(\lambda, N) \triangleq \ln E_{f\sim\pi}\mathbf{E}[e^{\lambda(\mathcal{L}(f) - \hat{\mathcal{L}}_N(f))}] \tag{9}$$

$\mathcal{M}_\pi$ is the set of all absolutely continuous densities w.r.t π, and $r_N(\lambda, \hat{\rho}, \delta)$ is the error term. That is, the PAC-Bayesian bound holds for every posterior $\hat{\rho}$ in $\mathcal{M}_\pi$, simultaneously.

We may think of π as a prior distribution density function and $\hat{\rho}$ as any candidate to a posterior distribution on the space of predictors. The inequality (7) says that the average generalization loss for models sampled from the posterior distribution is smaller than the average empirical loss for the posterior distribution plus the error terms r_N, with arbitrarily high probability.

A learning algorithm can be thought of as fixing a prior π and then choosing a posterior $\hat{\rho}$ for which the right-hand side of the inequality (7) is small. The latter can be viewed as a cost function involving the empirical loss and the regularization term r_N. The learned model is either sampled from the posterior density $\hat{\rho}$, or it is chosen as the one with maximal likelihood w.r.t. $\hat{\rho}$. Inequality (7) then gives guarantees on the generalization loss of the learned model. For more details on using PAC-Bayesian bounds see (Alquier, 2021).

The density which minimizes $\underset{f\sim\hat{\rho}}{E}\hat{\mathcal{L}}_N(f)(\omega) + r_N(\lambda, \hat{\rho}, \delta)$ is known as the Gibbs-posterior (Alquier, 2021) and it can be explicitly computed, i.e.

$$\rho_{\text{Gibbs}}(f) \triangleq Z^{-1}\pi(f)\exp(-\lambda\hat{\mathcal{L}}_N(f)), \qquad Z \triangleq E_{f\sim\pi}\exp(-\lambda\hat{\mathcal{L}}_N(f)). \tag{10}$$

In particular, using standard techniques (Alquier, 2021) it follows that

$$\underset{f\sim\rho_{Gibbs}}{E}\mathcal{L}(f) \leq \inf_{\hat{\rho}\in\mathcal{M}_\pi}\left(\underset{f\sim\hat{\rho}}{E}\hat{\mathcal{L}}_N(f) + r_N(\lambda, \hat{\rho}, \delta)\right)$$

with probability at least $1 - \delta$. One can also use PAC-Bayesian bounds, in order to choose the prior π or the hypothesis class $\mathcal{F}$, s.t. the difference between generalised

loss and empirical loss is within some acceptable level, i.e. $E_{f\sim\rho}\left(\mathcal{L}(f)-\hat{\mathcal{L}}_N(f)\right) \leq r_N(\lambda,\hat{\rho},\delta) \leq \epsilon$, after which one can proceed with more standard Bayesian learning approach on just the empirical loss $\hat{\mathcal{L}}_N(f)$.

From the discussion above it follows that it is desirable for the bound (7) to be tight. In particular, as the empirical loss converges to the generalization loss as $N\to\infty$, we expect that for tight bounds the term $r_N(\lambda,\hat{\rho},\delta)$ should converge to a small, preferably zero, constant as $N\to\infty$.

4. Main Results

In this paper we derive PAC-Bayesian bounds (7) for LTI systems. The main idea is to use the change of measure inequality from (Germain et al., 2016, Theorem 3). The major challenge is to bound the corresponding moment generating function/higher-order moments of $(\mathcal{L}(f)-\hat{\mathcal{L}}_N(f))$. However this brings some technical challenges. Namely, the processes involved are not i.i.d.. Moreover, they are not bounded, and the quadratic loss function is not Lipschitz. In addition, the empirical loss $\hat{\mathcal{L}}_N(f)$ is not an unbiased estimate of the generalization loss $\mathcal{L}(f)$. This is specific to state-space representations, for auto-regressive models considered in (Alquier & Wintenberger, 2012; Alquier et al., 2013; Alquier & Guedj, 2018) this problem does not occur. All these issues make it impossible to directly apply existing techniques (Alquier & Wintenberger, 2012; Alquier et al., 2013; Alquier & Guedj, 2018).
As the first step, we replace the empirical loss $\hat{\mathcal{L}}_N(f)$ by

$$V_N(f) \triangleq \frac{1}{N}\sum_{i=0}^{N-1}(\mathbf{y}(i)-\hat{\mathbf{y}}_f(i))^2 \tag{11}$$

where the finite-horizon prediction $\hat{\mathbf{y}}_f(t\mid 0)$ is replaced by the infinite horizon prediction $\hat{\mathbf{y}}_f(t)$ defined in Lemma 2.4. The advantage of $V_N(f)$ over $\hat{\mathcal{L}}_N(f)$ is that $V_N(f)$ is an unbiased estimate of the generalization loss $\mathcal{L}(f)$, i.e., $\mathbf{E}[V_N(f)]=\mathcal{L}(f)$. Indeed, since $\mathbf{y}(t)-\hat{\mathbf{y}}_f(t)$ is a stationary process, $\mathbf{E}[\|\mathbf{y}(i)-\hat{\mathbf{y}}_f(i)\|_2^2]=\mathcal{L}(f)$ does not depend on i, and hence $\mathbf{E}[V_N(f)]=\frac{1}{N}\sum_{i=0}^{N-1}\mathbf{E}[\|\mathbf{y}(i)-\hat{\mathbf{y}}_f(i)\|_2^2]=\mathcal{L}(f)$. Usual techniques for deriving error bounds are easier to extend to $V_N(f)$ than to $\hat{\mathcal{L}}_N(f)$. In order to derive upper bounds on the errors of the type (8), we will first derive upper bounds of the type (8), for $\mathcal{L}(f)-V_N(f)$, secondly we will derive upper bounds for $V_N(f)-\hat{\mathcal{L}}_N(f)$, then we will combine them using union bound. Doing this might seem counter-productive, however it is significantly easier to bound moments, $\mathbf{E}[(\mathcal{L}(f)-V_N(f))^r]$, and $\mathbf{E}[(V_N(f)-\hat{\mathcal{L}}_N(f))^r]$.

For every predictor f we define a number of constants which will be used in the PAC-Bayesian error bound. Let A_g,K_g,C_g be the matrices of the data generator from Assumption 2.1. Let us define the matrices (A_e,K_e,C_e,D_e) as $D_e=I-\hat{D}_w$,

$$A_e=\begin{bmatrix}A_g & 0\\ \hat{B}C_w & \hat{A}\end{bmatrix}, K_e=\begin{bmatrix}K_g\\ \hat{B}_w\end{bmatrix}, C_e=\begin{bmatrix}(C_1-\hat{D}C_w)^T\\ -\hat{C}^T\end{bmatrix}^T$$

where $C_g=\begin{bmatrix}C_1^T & C_2^T\end{bmatrix}^T$ and C_1 has $n_{\mathbf{y}}$ rows and C_2 has $n_{\mathbf{u}}$ rows; and

$$(C_w,\hat{B}_w,\hat{D}_w)=\begin{cases}(C_2,\begin{bmatrix}0 & \hat{B}\end{bmatrix},\begin{bmatrix}0 & \hat{D}\end{bmatrix}) & \text{if } \mathbf{w}=\mathbf{u}\,,\\ (C_g,\hat{B},\hat{D}) & \text{if } \mathbf{w}=\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\end{bmatrix}\end{cases}$$

The matrices A_e,K_e,C_e,D_e represent the LTI system driven by the innovation process $\mathbf{e}_g$ of $(\mathbf{y}^T,\mathbf{w}^T)^T$, output of which is $\mathbf{y}-\hat{\mathbf{y}}_f$, i.e.,

$$\begin{aligned}\tilde{\mathbf{x}}(t+1)&=A_e\tilde{\mathbf{x}}(t)+K_e\mathbf{e}_g(t),\\ \mathbf{y}(t)-\hat{\mathbf{y}}_f(t)&=C_e\tilde{\mathbf{x}}(t)+D_e\mathbf{e}_g(t)\end{aligned} \tag{12}$$

Next, choose $\hat{M}(f)>1$, and let $\hat{\gamma}(f)\in[\hat{\gamma}^*(f),1)$, such that $\|\hat{A}^k\|_2\leq\hat{M}(f)\hat{\gamma}^k(f)$, where $\hat{\gamma}^*(\hat{A})$ is the spectral radius of $\hat{A}$.

Definition 4.1 (Constants $\bar{G}_f(f),G_e(f)$)**.** With these definitions,

$$G_e(f)\triangleq\|(A_e,K_e,C_e,D_e)\|_{\ell_1}\triangleq\|D_e\|_2+\sum_{k=0}^{\infty}\|C_eA_e^kK_e\|_2$$

$$G_{e,1}(f)\triangleq\|D_e\|_2+\sum_{k=0}^{\infty}(k+1)\|C_eA_e^kK_e\|_2$$

$$\|\Sigma_{gen}\|_{\ell_1}\triangleq 1+\sum_{k=0}^{\infty}\|C_gA_g^{k-1}K_g\|_2$$

$$\bar{G}_f(f)\triangleq\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{(1-\hat{\gamma})^{1.5}}$$

The interpretation of the various terms appearing in Definition 4.1 is as follows.

Remark 4.2 (Interpretation of constants).
The term $G_e(f)$ is the ℓ_1 norm of the error system, it relates high-order moments of the infinite past prediction error and the high-order moments of the innovation noise $\mathbf{e}_g$ of the data generator, i.e., $\mathbf{E}[\|\mathbf{y}(t)-\hat{\mathbf{y}}(t)\|_2^r]\leq G_e^r(f)\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]$.
The term $G_{e,1}(f)$ characterizes mixing properties (short memory condition) (Alquier et al., 2013) of the prediction error $\mathbf{y}(t)-\hat{\mathbf{y}}(t)$. Intuitively,the smaller $G_{e,1}(f)$ is, the more the prediction error behaves like an i.i.d process. This constant is related to stability of the error system, i.e., the smaller the spectral radius of A_e is, the smaller this constant is.
The term $\|\Sigma_{gen}\|_{\ell_1}$ is the ℓ_1 norm of the data generator, it relates high-order moments of $\mathbf{y}$ and $\mathbf{u}$ with those of the

noise, i.e., $\mathbf{E}[\|[\mathbf{y}^T(t),\mathbf{u}^T(t)]^T\|_2^r] \le \|\Sigma_{gen}\|_{\ell_1}^r \mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]$. **The term** $\bar{G}_f(f)$ depends only the predictor f, and it gives an upper bound on high-order moments of the difference of empirical losses $V_N(f) - \mathcal{L}_N(f)$ in terms of high-order moments of $\mathbf{y}$ and $\mathbf{u}$, i.e, it is shown in (Eringis et al., 2023) that $\mathbf{E}[|V_N(f) - \mathcal{L}_N(f)|^r\| \le \frac{2^r}{\sqrt{N}}\bar{G}_f(f)^r\mathbf{E}[\|[\mathbf{y}^T(t),\mathbf{u}^T(t)]^T\|^r]$. This terms decreases with the spectral radius of the predictor, the smaller is $\bar{G}_f(f)$.

Theorem 4.3. *Let $\mathcal{M}_\pi$ denote the set of all absolutely continuous densities w.r.t π. Let $\mu_{max}(Q_e)$ be the maximal eigenvalue of the covariance matrix $Q_e = E[\mathbf{e}_g(t)\mathbf{e}_g^T(t)]$ of the noise $\mathbf{e}_g$ from Assumption 2.1. Then for any density π on hypothesis class $\mathcal{F}$, any $\delta \in (0,1]$, and*

$$0 < \lambda < \Big(\sup_{f\in\mathcal{F}} \max\{8(n_{\mathbf{y}}+n_{\mathbf{u}})\bar{G}_{gen}\bar{G}_f(f), \\ 6(n_{\mathbf{y}}+n_{\mathbf{u}}+1)n_{\mathbf{y}}\mu_{\max}(Q_e)G_e(f)^2\}\Big)^{-1} \quad (13)$$

the following inequality holds with probability at least $1-2\delta$

$$\forall \rho \in \mathcal{M}_\pi : E_{f\sim\hat{\rho}}\mathcal{L}(f) \le \mathbb{E}_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + r_N(\lambda,N), \quad (14)$$

with

$$r_N(\lambda,\hat{\rho},\delta) \triangleq \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \widehat{\Psi}_\pi(\lambda,N)\right]$$

$$\widehat{\Psi}_\pi(\lambda,N) \triangleq \frac{1}{2}\left(\ln \underset{f\sim\pi}{E}\widehat{\Psi}_1(f) + \ln \underset{f\sim\pi}{E}\widehat{\Psi}_2(f)\right)$$

$$\widehat{\Psi}_1(f) \triangleq 1 + \frac{2(m+1)!\left(6\lambda n_{\mathbf{y}}\mu_{\max}(Q_e)G_e(f)^2\right)^2}{N(1-6(m+1)\lambda n_{\mathbf{y}}\mu_{\max}(Q_e)G_e(f)^2)}$$

$$\widehat{\Psi}_2(f) \triangleq 1 + \frac{8(m!)\lambda\bar{G}_f(f)\|\Sigma_{gen}\|_{\ell_1}^2\mu_{\max}(Q_e)}{\sqrt{N}\left(1-8\lambda m\bar{G}_f(f)\|\Sigma_{gen}\|_{\ell_1}^2\mu_{\max}(Q_e)\right)}$$

Sketch of the proof of Theorem 4.3. Below we will skecth the basic steps, for a detailed proof of Theorem 4.3, see (Eringis et al., 2023, Proof A.17). The main idea of the proof is to show that for $i = 1,2$

$$\forall\rho\in\mathcal{M}_\pi : E_{f\sim\hat{\rho}}Y_i(f) \le E_{f\sim\hat{\rho}}X_i(f) + \\ \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \ln E_{f\sim\pi}\widehat{\Psi}_{N,i}(0.5\lambda,f)\right],$$

with probability $1-\delta$, where $Y_1(f) = \mathcal{L}(f)$, $Y_2(f) = X_1(f) = V_N(f)$, $X_2(f) = \hat{\mathcal{L}}_N(f)$, and then combine these two inequalities to derive (14). In order to derive these two inequalities, we use the Donsker-Varadhan change of measure inequality applied to $Y_i(f) - X_i(f)$, $i = 1,2$ and the following inequalities on the moment generating function of $\mathcal{L}(f) - V_N(f)$ and $V_N(f) - \mathcal{L}_N(f)$

$$\mathbf{E}\left[e^{\lambda(\mathcal{L}(f)-V_N(f))}\right] \le \widehat{\Psi}_{N,1}(0.5\lambda,f) \quad (15)$$

$$\mathbf{E}[e^{\lambda(V_N(f)-\hat{\mathcal{L}}_N(f))}] \le \widehat{\Psi}_{N,2}(0.5\lambda,f) \quad (16)$$

In order to prove (15) – (16), we use the following bounds on the high-order moments of $\mathcal{L}(f) - V_N(f)$ and $V_N(f) - \mathcal{L}_N(f)$:

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \le \frac{\sigma(r)}{N}4(r-1)n_{\mathbf{y}}^r G_e(f)^{2r} \quad (17)$$

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \le \frac{\bar{\sigma}(r)}{\sqrt{N}}\left(4\bar{G}_f(f)\|\Sigma_{gen}\|_{\ell_1}^2\right)^r \quad (18)$$

where $\sigma(r)$ and $\bar{\sigma}(r)$ satisfies

$$\sigma(r) = \sup_{t,k,j}\mathbf{E}\left[\|\mathbf{e}(t,k,j)\|_2^r\right],$$

$$\mathbf{e}(t,k,j) \triangleq \mathbf{E}[\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j)] - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j)$$

$$\bar{\sigma}(r) \ge \mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]$$

If $\mathbf{e}_g(t)$ is sub-Gaussian, then $\sigma(r) \le 3^r\mu_{\max}(Q_e)^r(m+r-1)!$, and $\bar{\sigma}(r) \le 2^r\mu_{\max}(Q_e)^r(m+r-1)!$. Using this observation after some manipulation, we can derive from (17)-(18) the inequalities (15)-(16) □

Discussion on the error bound The error bound of Theorem 4.3 is relatively intuitive: it is increasing with the noise covariance of the generator system, and with the ℓ_1 norm of the generator system, the error system, and the predictor. The higher the noise covariance is, the noisier is the generated data, and the more difficult it is to predict it. The higher the various norms and constants are, the less the generator, error system and the predictors are able to suppress the effects of the noise. In particular, the smaller is the spectral radius of the predictors and of the generator, the smaller is the error bound.

Note that, as $N \to \infty$ the PAC-Bayesian error term $r_N(\lambda,\hat{\rho},\delta)$ converges to the constant $\frac{1}{\lambda}\left(D_{\mathrm{KL}}(\hat{\rho}|\pi) + \ln\left(\frac{1}{\delta}\right)\right)$ for any posterior $\hat{\rho}$ and $\lambda > 0$. That is, irrespective of $\hat{\rho},\pi$, the error $r_N \ge \frac{1}{\lambda}\ln\left(\frac{1}{\delta}\right)$. Usually, one chooses $\lambda = \lambda_N$ as an increasing function of N, which then allows the PAC-Bayesian error $r_N(\lambda_N,\hat{\rho},\delta)$ to converge to 0 for any posterior $\hat{\rho}$ and confidence level $\delta > 0$. However, since by Theorem 4.3, λ is bounded by a constant, we can not control the term $\frac{1}{\lambda}\ln\left(\frac{1}{\delta}\right)$, and hence $r_N(\lambda,\hat{\rho},\delta)$ will not converge to zero.

A similar problem was already observed for PAC-Bayesian bounds for linear regression with unbounded signals and Gaussian noise (Shalaeva et al., 2020). In the next section, we eliminate this problem for the case when the noise $\mathbf{e}_g$ of the data generator is bounded.

Remark 4.4 (Relationship with bounds from (Shalaeva et al., 2020; Eringis et al., 2021)). Note that the bound above not only applies to a more general class of models than (Eringis et al., 2021; Shalaeva et al., 2020), but it is also

asymptotically tighter that the bounds of (Eringis et al., 2021; Shalaeva et al., 2020). Indeed, the latter bounds converge to $\frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}|\pi) + \ln\left(\frac{1}{\delta}\right)\right] + c$ as $N \to \infty$ for some constant $c > 0$, while the bound of Theorem 4.3 converges to the same expression but with $c = 0$.

5. Bounded case

As it was mentioned above, for the case of bounded signals Theorem 4.3 can be improved. More precisely, in this section we will use the following assumption.

Assumption 5.1. The noise $\mathbf{e}_g(t)$ from Assumption 2.1 is essentially bounded, i.e. with probability 1, $\max_{i=1,\dots,m} |\mathbf{e}_{g,i}(t)| \leq c_e$, for some $c_e > 0$. Moreover, $\|\mathbf{e}_g(t)\|_2 \leq C \triangleq c_e\sqrt{m}$

With the assumption above $\mathbf{y}(t)$ and $\mathbf{u}(t)$ are essentially bounded for all $t \in \mathbb{Z}$. This then allows us to derive the following sharper bounds.

Theorem 5.2. *Let $\mathcal{M}_\pi$ denote the set of all absolutely continuous densities w.r.t π. Under assumption 5.1 it holds true that for any density π on hypothesis class $\mathcal{F}$, any $\delta \in (0, 1]$, and $\lambda > 0$ the following inequality holds with probability at least $1 - 2\delta$*

$$\forall \hat{\rho} \in \mathcal{M}_\pi : E_{f\sim\hat{\rho}}\mathcal{L}(f) \leq E_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + \bar{r}_N(\lambda, \hat{\rho}, \delta) \tag{19}$$

with the constants

$$\begin{aligned}
\bar{r}_N(\lambda, \hat{\rho}, \delta) &\triangleq \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}||\pi) + \ln\frac{1}{\delta} + \bar{\Psi}_\pi(\lambda, N)\right] \\
\bar{\Psi}_\pi(\lambda, N) &\triangleq \frac{1}{2}\left(\ln \underset{f\sim\pi}{E}\bar{\Psi}_1(f) + \ln \underset{f\sim\pi}{E}\bar{\Psi}_2(f)\right) \\
\bar{\Psi}_1(f) &\triangleq 1 + \frac{1}{N}e^{\lambda 8 n_{\mathbf{y}} C^2 G_e(f)^2} \\
\bar{\Psi}_2(f) &\triangleq 1 + \frac{1}{\sqrt{N}}e^{\lambda 4\|\Sigma_{gen}\|_{\ell_1}^2 C^2 \bar{G}_f(f)}
\end{aligned}$$

Sketch of the proof of Theorem 5.2. The proof is similar to that of Theorem 5.2, but $\sigma(r)$ and $\bar{\sigma}(r)$ can be taken as $(2C)^r$ and C^r respectively, for a detailed proof see (Eringis et al., 2023, Corollary A.3). □

Discussion on the bound The intuitive interpretation of the bound of Theorem 5.2 in terms of the effect of the noise levels of the data generator, various system norms and stability is analogous to that of Theorem 4.3. However, in contrast to Theorem 4.3, λ is not bounded in Theorem 5.2, and as such we can choose $\lambda = \lambda_N$ an increasing function of N, in order to control the term $\frac{1}{\lambda_N}\ln\delta^{-1}$:

Corollary 5.3 (Bounds converging to zero)**.** *With the assumptions and notation of Theorem 5.2, then for any given posterior $\hat{\rho}$, $\lim_{N\to\infty}\bar{r}_N(\lambda_N, \hat{\rho}, \delta) = 0$, where*

$$\lambda_N = \frac{\ln\sqrt{N}}{C^2 \sup_{f\in\mathcal{F}} \max\{8n_{\mathbf{y}}G_e(f)^2, 4\|\Sigma_{gen}\|_{\ell_1}^2 \bar{G}_f(f)\}},$$

If one chooses λ_N as in Corollary 5.3, and considers posteriors $\hat{\rho}_N$ which depend on N, for example Gibbs posteriors (10), then it is hard to say what will happen with $\lambda_N^{-1}(D_{\mathrm{KL}}(\hat{\rho}_N\|\pi)$. This is a general issue with PAC-Bayesian bounds. However, simulations indicate that if λ_N is any reasonable increasing function of N, then $\lambda_N^{-1}D_{\mathrm{KL}}(\hat{\rho}_N\|\pi)$, will converge to some problem dependant constant.

The bound above has all the desired properties, but its rate of convergence to zero as $N \to +\infty$ is very slow, it is $O(\frac{1}{\ln\sqrt{N}})$. In fact, using (Alquier et al., 2013), the results of Theorem 5.2 can be sharpened as follows.

Theorem 5.4. *Let $\mathcal{M}_\pi$ denote the set of all absolutely continuous densities w.r.t π. Under assumption 5.1, for any density π on hypothesis class $\mathcal{F}$, any $\delta \in (0, 1]$, and $\lambda > 0$ the following inequality holds with probability at least $1 - 2\delta$,*

$$\forall \hat{\rho} \in \mathcal{M}_\pi : E_{f\sim\hat{\rho}}\mathcal{L}(f) \leq E_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + \tilde{r}_N(\lambda, \hat{\rho}, \delta) \tag{20}$$

where

$$\begin{aligned}
\tilde{r}_N(\lambda, \hat{\rho}, \delta) &\triangleq \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}||\pi) + \ln\frac{1}{\delta} + \tilde{\Psi}_\pi(\lambda, N)\right] \\
\tilde{\Psi}_\pi(\lambda, N) &\triangleq \frac{1}{2}\left(\ln \underset{f\sim\pi}{E}\tilde{\Psi}_1(f) + \ln \underset{f\sim\pi}{E}\tilde{\Psi}_2(f)\right) \\
\tilde{\Psi}_1(f) &\triangleq \left(1 - C_{1,2}(f) + C_{1,2}(f)e^{\frac{\lambda}{N}2C_{1,1}(f)}\right) \\
\tilde{\Psi}_2(f) &\triangleq \left(e^{\frac{\lambda^2}{N}8(G_e(f)+G_{e,1}(f))^2C^2(4G_e(f)C+1)^2}\right) \\
C_{1,i}(f) &\triangleq \bar{G}_{f,i}(f)\|\Sigma_{gen}\|_{\ell_1}C \\
\bar{G}_{f,1}(f) &\triangleq \hat{M}\|\hat{C}\|_2\|\hat{B}\|_2\left(1-\hat{\gamma}\right)^{-1}, \\
\bar{G}_{f,2}(f) &\triangleq \left(1 + \|\hat{D}\|_2 + \bar{G}_{f,1}(f)\right)\left(1-\hat{\gamma}\right)^{-1}
\end{aligned}$$

Sketch of the proof of Theorem 5.4. We repeat the steps of the proof of Theorem 4.3. However, we replace (15) by $\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}] \leq \tilde{\Psi}_{N,1}(0.5\lambda, f)$ which is derived using the extension of Hoefding's inequality in (Alquier et al., 2013, Theorem 6.6). We then replace (16) with $\mathbf{E}[e^{\lambda|V_N(f)-\hat{\mathcal{L}}_N(f)|}] \leq \tilde{\Psi}_{N,2}(0.5\lambda, f)$. The latter follows by using the bounds $\bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}C \cdot \left(\frac{2\|\Sigma_{gen}\|_{\ell_1}C}{N}\bar{G}_{f,2}(f)\right)^r$ on the moments $\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r]$ instead of (18). For the detailed proof see (Eringis et al., 2023, Proof A.26). □

Discussion on the bound The intuition behind the role of the noise level, system norms and stability for the bound above is the same as for the bounds of the previous theorems. The bound above has the advantage that it converges to zero significantly faster than the convergence in Corollary 5.3:

Corollary 5.5 (Fast convergence $O(\frac{1}{\sqrt{N}})$**.** *If* $\lambda_N = \sqrt{N}$*, then for any given posterior* $\hat{\rho}$*,* $\lim_{N\to\infty} \tilde{r}_N(\lambda_N, \hat{\rho}, \delta) = 0$ *with the rate of convergence* $O(\frac{1}{\sqrt{N}})$*.*

As before, if the posterior $\hat{\rho}_N$ depends on N, say, it is the Gibbs posterior, then it is difficult to prove convergence $\tilde{r}_N(\lambda_N, \hat{\rho}_N, \delta)$ analytically. This is a general issue with PAC-Bayesian bounds. However, if $D_{\mathrm{KL}}(\hat{\rho}_N|\pi)$ grows slower than $\frac{1}{\sqrt{N}}$, then the error bound $\tilde{r}_N(\lambda_N, \hat{\rho}_N, \delta)$ will still converge to zero as $N \to \infty$ for $\lambda_N = \sqrt{N}$. Numerical simulations reveal, for the Gibbs posterior this is the case.

6. Numerical example

For the sake of illustration let us assume that data is generated by

$$\mathbf{x}(t+1) = \begin{bmatrix} 0.16 & -0.3 \\ 0 & -0.05 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0.33 & -0.75 \\ 0 & -0.09 \end{bmatrix} \mathbf{e}_g(t)$$
$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{x}(t) + \mathbf{e}_g(t),$$

Following the two theorems in the paper, we will consider two cases:
(1) Unbounded innovation noise: $\mathbf{e}_g(t) \sim \mathcal{N}(0, Q_e)$, $Q_e = \begin{bmatrix} 0.054 & 0.018 \\ 0.018 & 0.248 \end{bmatrix}$,
(2) Bounded innovation noise: $\mathbf{e}_g(t)$ is distributed according to zero-mean truncated gaussian, s.t. $c_e = 1$, and $\mathbf{E}[\mathbf{e}_g(t)\mathbf{e}_g^T(t)] \approx Q_e$.

We will assume that the predictors are fully parametrised, i.e all entries of matrices $\hat{A}(\theta), \hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta)$ are parametrised, and all predictors are second-order systems, i.e. $\hat{A}(\theta) \in \mathbb{R}^{2\times 2}$. Moreover in the case of $\mathbf{w}(t) = [\mathbf{y}^T(t), \mathbf{u}^T(t)]^T$, we take $\hat{D}(\theta) = \begin{bmatrix} 0 & \theta_9 \end{bmatrix}$. Thus, with $\Sigma(\theta) = (\hat{A}(\theta), \hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta))$, we will define our hypothesis class to be

$$\mathcal{F} = \{f_{\Sigma(\theta)} | \gamma(\hat{A}(\theta)) < 1, \bar{G}_f(f_{\Sigma(\theta)}) < 10, \theta \in \mathbb{R}^{11}\}$$

We shall use the prior, $\pi(f) = Z_\pi \exp(-\bar{G}_f(f))$, with Z_π the normalisation term. This prior will act as regularisation, penalising predictors with high ℓ_1 norms. This in turn will reduce the term $E_{f\sim\pi}(1 + \frac{1}{\sqrt{N}} e^{\lambda G_{gen,2} \bar{G}_f(f)})$. We will use the Gibbs posterior $\rho_N(f) = Z_\rho \pi(f) \exp(-\lambda(N)\hat{\mathcal{L}}_N(f))$ In order to compute the numerical value of r_N, we can use Markov-Chain Monte-Carlo methods, which means that we only need to be able to evaluate $\hat{\pi}(f) \propto \pi(f)$, and $\hat{\rho}(f) \propto \rho(f)$. More precisely one can approximate r_N, by only being able to evaluate $\hat{\pi}(f)$ and $\beta(f) \triangleq \frac{\hat{\rho}(f)}{\hat{\pi}(f)} \propto \frac{\rho(f)}{\pi(f)}$

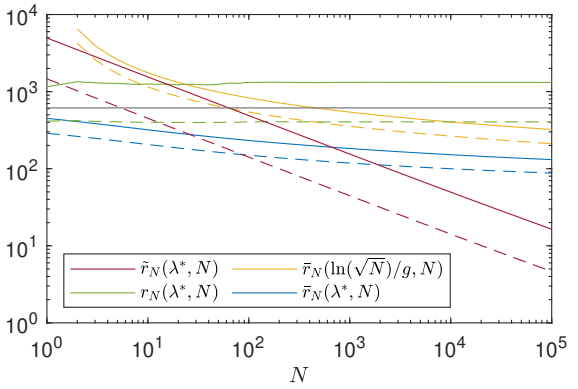

Figure 1. Numerical simulation of both cases (bounded and unbounded noise), solid lines depict case of $\mathbf{w} = \mathbf{u}$, dashed lines show case of $\mathbf{w} = [\mathbf{y}^T, \mathbf{u}^T]^T$, λ^* is found by numerical optimisation, i.e. $\lambda^* = \arg\min_\lambda r_N(\lambda, N)$, the black horizontal line denotes a vacuous bound for the bounded noise case, i.e. any bounds above that line are vacuous

In Fig. 1 we see the convergence of the error terms, for the case of bounded noise. Note that the proposed function λ_N is close to numerically optimal (blue line in Fig. 1), i.e. asymptotically $\lambda_N \propto \ln\sqrt{N}$, seem to be optimal, one could try to find a less conservative scale of λ_N. Note that, in this example, for $N \leq 460$, Theorem 5.2, yields vacuous bounds, i.e. $\bar{r}_N(\frac{\ln\sqrt{N}}{g}, \hat{\rho}, \delta) \geq 2(C \sup_{f\in\mathcal{F}} G_e(f))^2$, the highest possible error. However for Theorem 5.4, only for $N \leq 64$, is the bound vacuous.

For the case of unbounded innovation noise, as stated before we see in Fig. 1 that it converges to a constant. Unfortunately, since λ is bounded not much can be done. However, since the noise is unbounded it is difficult to determine if the bound is vacuous.

7. Conclusion

In this paper we have derived PAC-Bayesian error bounds for stochastic LTI systems with inputs. For data generated by an LTI system with sub-gaussian noise, the error bound is asymptotically bounded from below, which indicates that the bound is not tight. For data generated by an LTI system with bounded innovation noise, the error bound converges to zero at the rate $O(\frac{1}{\sqrt{N}})$, which is comparable to most of PAC-Bayesian bounds.

Future research will be directed towards extending these results to more general state-space representations and using the results of the paper for deriving oracle inequalities (Alquier, 2021).

8. Acknowledgements

This work was partially funded by the French government's Future Investments program within the framework of the SystemX Technological Research Institute and Confiance.ia initiative, and by the IEA program of CNRS.

References

Alquier, P. User-friendly introduction to pac-bayes bounds. *arXiv:2110.11216*, 2021.

Alquier, P. and Guedj, B. Simpler PAC-Bayesian Bounds for Hostile Data. *Machine Learning*, 107(5):887–902, 2018.

Alquier, P. and Wintenberger, O. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883 – 913, 2012.

Alquier, P., Li, X., and Wintenberger, O. Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 1(2013):65–93, 2013.

Caines, P. E. *Linear Stochastic Systems*. John Wiley and Sons, 1988.

Campi, M. C. and Weyer, E. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

Chen, M., Li, X., and Zhao, T. On generalization bounds of a family of recurrent neural networks. In *Proceedings of AISTATS 2020*, volume 108 of *PMLR*, pp. 1233–1243, 8 2020.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*. AUAI Press, 2017.

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., Esfahan, A. F., and Petreczky, M. Pac-bayesian theory for stochastic lti systems. In *2021 60th IEEE CDC*, pp. 6626–6633, 2021.

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., and Petreczky, M. Explicit construction of the minimum error variance estimator for stochastic lti state-space systems. *arXiv:2109.02384*, 2022.

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., and Petreczky, M. Pac-bayesian bounds for learning lti-ss systems with input from empirical loss. *arXiv:2303.16816*, 2023.

Foster, D. and Simchowitz, M. Logarithmic regret for adversarial online control. In *Proceedings of the 37th ICML*, volume 119 of *PMLR*, pp. 3211–3221. PMLR, 7 2020.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. Pac-bayesian theory meets bayesian inference. In *NIPS*, pp. 1876–1884, 2016.

Guedj, B. A Primer on PAC-Bayesian Learning. *arXiv:1901.05353*, 2019.

Hannan, E. and Deistler, M. *The Statistical Theory of Linear Systems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1988.

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Joukovsky, B., Mukherjee, T., Van Luong, H., and Deligiannis, N. Generalization error bounds for deep unfolding rnns. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *PMLR*, pp. 1515–1524. PMLR, 7 2021.

Koiran, P. and Sontag, E. D. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

Lindquist, A. and Picci, G. *Linear Stochastic Systems: A Geometric Approach to Modeling, Estimation and Identification*. Springer, 2015.

Ljung, L. *System Identification: Theory for the user (2nd Ed.)*. PTR Prentice Hall., Upper Saddle River, USA, 1999.

Massucci, L., Lauer, F., and Gilson, M. Regularized switched system identification: a statistical learning perspective. *IFAC-PapersOnLine*, 54(5):55–60, 2021. ISSN 2405-8963. 7th IFAC Conference on Analysis and Design of Hybrid Systems ADHS 2021.

Oymak, S. and Ozay, N. Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4): 1914–1928, 4 2022.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite time LTI system identification. *J. Mach. Learn. Res.*, 22:26:1–26:61, 2021.

Shalaeva, V., Esfahani, A. F., Germain, P., and Petreczky, M. Improved PAC-bayesian bounds for linear regression. *Proceedings of the AAAI Conference*, 34:5660–5667, 4 2020.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Simchowitz, M. *Statistical Complexity and Regret in Linear Control*. University of California, Berkeley, 2021.

Simchowitz, M., Boczar, R., and Recht, B. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pp. 2714–2802. PMLR, 2019.

Sontag, E. D. A learning result for continuous-time recurrent neural networks. *Systems & control letters*, 34(3):151–158, 1998.

Steele, J. M. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Tsiamis, A. and Pappas, G. J. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3648–3654, 2019.

Vidyasagar, M. and Karandikar, R. L. A learning theory approach to system identification and stochastic adaptive control. *Probabilistic and randomized methods for design under uncertainty*, pp. 265–302, 2006.

Zhang, T. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Information Theory*, 52(4):1307–1321, 2006.

A. Proofs

In this section we provide the proofs of theorem 4.3 and 5.2 under the assumptions stated in the main text. To do so we first prove a series of lemmas.

Lemma A.1. *For random variable $\mathbf{e}_g(t) \sim \mathcal{N}(0, Q_e)$, the following holds*

$$\begin{aligned}\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r] &\leq \mu_{\max}(Q_e)^{\frac{r}{2}}\mathbf{E}[\|\mathbf{z}(t)\|_2^r]\\ \mathbf{z}(t) &\sim \mathcal{N}(0, I),\end{aligned}$$

where $Q_e = \mathbf{E}[\mathbf{e}_g(t)\mathbf{e}_g^T(t)]$, and $\mu_{\max}(Q_e)$ denotes the maximal eigen value of Q_e.

Proof A.1 (Proof of Lemma A.1). First, note $\mathbf{z}(t) = Q_e^{-\frac{1}{2}}\mathbf{e}_g(t)$, and

$$\|\mathbf{e}_g(t)\|_2^2 = \mathbf{e}_g^T(t)\mathbf{e}_g(t) = \mathbf{z}^T(t)Q_e^{\frac{1}{2}}Q_e^{\frac{1}{2}}\mathbf{z}(t) = \mathbf{z}^T(t)Q_e\mathbf{z}(t)$$

therefore

$$\begin{aligned}\|\mathbf{e}_g(t)\|_2^2 &\leq \mu_{\max}(Q_e)\|\mathbf{z}(t)\|_2^2\\ \|\mathbf{e}_g(t)\|_2^r &\leq \mu_{\max}(Q_e)^{\frac{r}{2}}\|\mathbf{z}(t)\|_2^r\\ \mathbf{E}[\|\mathbf{e}_g(t)\|_2^r] &\leq \mu_{\max}(Q_e)^{\frac{r}{2}}\mathbf{E}[\|\mathbf{z}(t)\|_2^r]\end{aligned}$$

Finally, note that $\mathbf{z}(t) \sim \mathcal{N}(0, I)$.

Lemma A.2. *If $\mathbf{z}(t) \sim \mathcal{N}(0, I_m)$, then*

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^r]^2 \leq 4((m + r - 1)!)$$

Proof A.2 (Proof of Lemma A.2). First, notice that the distribution of $\|\mathbf{z}(t)\|_2 = \sqrt{\sum_{i=1}^m \mathbf{z}_i^2(t)}$ is chi- distribution, as such

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^r] = 2^{\frac{r}{2}}\frac{\Gamma(\frac{m+r}{2})}{\Gamma(\frac{m}{2})} \tag{21}$$

We will use mathematical induction to prove the lemma.
For $r = 0$, lemma holds, since

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^0]^2 = \left(2^{\frac{0}{2}}\frac{\Gamma(\frac{m+0}{2})}{\Gamma(\frac{m}{2})}\right)^2 = 1 \leq 4(m-1)!, \quad \forall m \in \mathbb{N}. \tag{22}$$

for $r = 1$, lemma holds, as

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^1] = 2^{\frac{1}{2}}\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})}.$$

Notice that, for scalar $\mathbf{x} \sim \mathcal{N}(0, 1)$

$$\mathbf{E}[|\mathbf{x}|^k] = 2^{\frac{k}{2}}\frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi}}$$

It is also known that

$$\mathbf{E}[|\mathbf{x}|^k] = \begin{cases}(k-1)!!\sqrt{\frac{2}{\pi}}, & k \text{ odd}\\ (k-1)!!, & k \text{ even}\end{cases}$$

therefore,

$$2^{\frac{k}{2}}\frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi}} = \begin{cases}(k-1)!!\sqrt{\frac{2}{\pi}}, & k \text{ odd}\\ (k-1)!!, & k \text{ even}\end{cases}$$

Applying this to $k = m$ and $k = m - 1$, we obtain

$$2^{\frac{m}{2}} \frac{\Gamma(\frac{m+1}{2})}{\sqrt{\pi}} = \begin{cases} (m-1)!!\sqrt{\frac{2}{\pi}}, & m \text{ odd} \\ (m-1)!!, & m \text{ even} \end{cases}$$

$$2^{\frac{m-1}{2}} \frac{\Gamma(\frac{m}{2})}{\sqrt{\pi}} = \begin{cases} (m-2)!!\sqrt{\frac{2}{\pi}}, & (m-1) \text{ odd}, (m \text{ even}) \\ (m-2)!!, & (m-1) \text{ even}, (m \text{ odd}) \end{cases}$$

Now notice,

$$\mathbf{E}[\|z(t)\|_2^1] = 2^{\frac{1}{2}} \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})} = \frac{2^{\frac{m}{2}} \frac{\Gamma(\frac{m+1}{2})}{\sqrt{\pi}}}{2^{\frac{m-1}{2}} \frac{\Gamma(\frac{m}{2})}{\sqrt{\pi}}} = \frac{(m-1)!!}{(m-2)!!} c_m$$

$$c_m = \begin{cases} \sqrt{\frac{2}{\pi}}, & m \text{ even} \\ \sqrt{\frac{\pi}{2}}, & m \text{ odd} \end{cases}$$

notice that $c_m \leq 2$ for all m, and therefore

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^1] \leq 2 \frac{(m-1)!!}{(m-2)!!} \leq 2(m-1)!! \tag{23}$$

Then

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^1]^2 \leq 4((m-1)!!)^2$$

Note that $((m-1)!!)^2 \leq m!$. We can see that by contradiction: assume that $((m-1)!!)^2 \geq m!$. Notice that $m! = m!!(m-1)!!$ and hence $((m-1)!!)^2 \geq m!$ implies $(m-1)!! \geq m!!$. As $(m-1)!!$ must be less than $m!!$ we have a contradiction. Therefore $((m-1)!!)^2 \leq m!$ holds and we have

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^1]^2 \leq 4m!.$$

That is, we have shown that for $r = 0$ and $r = 1$ Lemma A.2 holds.
Now suppose that for all $k \geq 2$ and for all $0 \leq r \leq k$

$$2^{\frac{r}{2}} \frac{\Gamma(\frac{m+r}{2})}{\Gamma(\frac{m}{2})} \leq 4(m+r-1)!, \tag{24}$$

We will show that (24) holds for $r = k + 1$ too. To this end, notice that

$$\Gamma\left(\frac{m+k}{2}\right) = \Gamma\left(\frac{m+k-2}{2} + 1\right) = \frac{m+k-2}{2} \Gamma\left(\frac{m+k-2}{2}\right)$$

Using this relation we obtain

$$\begin{aligned} \left(2^{\frac{k}{2}} \frac{\Gamma(\frac{m+k}{2})}{\Gamma(\frac{m}{2})}\right)^2 &= \left(\left(2^{\frac{k-2}{2}} \frac{\Gamma(\frac{m+k-2}{2})}{\Gamma(\frac{m}{2})}\right)\left(2\frac{m+k-2}{2}\right)\right)^2 \\ &= \left(2^{\frac{k-2}{2}} \frac{\Gamma(\frac{m+k-2}{2})}{\Gamma(\frac{m}{2})}\right)^2 \left(2\frac{m+k-2}{2}\right)^2. \end{aligned} \tag{25}$$

Now $k - 2 \in [0, k]$, so we can apply to it the induction hypothesis. That is, for $r = k - 2$, (24) holds, i.e.,

$$\left(2^{\frac{r}{2}} \frac{\Gamma(\frac{m+r}{2})}{\Gamma(\frac{m}{2})}\right) \leq 4(m+r-1)! = 4(m+k-3)!.$$

and therefore

$$\left(2^{\frac{k}{2}}\frac{\Gamma(\frac{m+k}{2})}{\Gamma(\frac{m}{2})}\right)^2 \leq 4(m+k-3)!\left(4\frac{(m+k-2)^2}{4}\right) = 4(m+k-3)!(m+k-2)(m+k-2).$$

Using $(m+k-2) \leq (m+k-1)$, it follows that

$$\left(2^{\frac{k-2}{2}}\frac{\Gamma(\frac{m+k-2}{2})}{\Gamma(\frac{m}{2})}\right)^2 \left(2\frac{m+k-2}{2}\right)^2 \leq 4(m+k-3)!(m+k-2)(m+k-2) \leq 4(m+k-1)!$$

Substituting the last inequality into (25), it follows that (24) holds for $r = k+1$.

Lemma A.3. *For random variable $\mathbf{z} \sim \mathcal{N}(0, I_m)$, the even moments of $\|\mathbf{z}\|_2$ are bounded by*

$$\mathbf{E}[\|\mathbf{z}\|_2^{2r}] \leq 2^r(m+r-1)!$$

Proof A.3 (Proof of Lemma A.3). Clearly $\|\mathbf{z}\|_2$ has the chi distribution,

$$\mathbf{E}[\|\mathbf{z}\|_2^{2r}] = 2^{\frac{2r}{2}}\frac{\Gamma(\frac{m+2r}{2})}{\Gamma(\frac{m}{2})} = 2^r\frac{\Gamma(\frac{m}{2}+r)}{\Gamma(\frac{m}{2})}$$

$$\Gamma\left(\frac{m}{2}+r\right) = \Gamma\left(\frac{m}{2}+(r-1)+1\right) = \left(\frac{m}{2}+(r-1)\right)\Gamma\left(\frac{m}{2}+(r-1)\right) = \left(\frac{m}{2}+(r-1)\right)\left(\frac{m}{2}+(r-2)\right)\ldots\frac{m}{2}\Gamma\left(\frac{m}{2}\right)$$

$$\mathbf{E}[\|\mathbf{z}\|_2^{2r}] = 2^r\frac{\left(\frac{m}{2}+(r-1)\right)\left(\frac{m}{2}+(r-2)\right)\ldots\frac{m}{2}\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}$$

notice $\frac{m}{2} \leq m$, then

$$\mathbf{E}[\|\mathbf{z}\|_2^{2r}] \leq 2^r\frac{(m+r-1)!}{m!} \leq 2^r(m+r-1)!$$

Combining Lemmas (A.1 and A.2), we obtain the following lemma.

Lemma A.4. *Let $r \in \mathbb{N}$*

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^{2r}] \leq \mu_{\max}(Q_e)^r 2^r(m+r-1)!$$

Combining Lemmas (A.1 and A.3), we obtain the following lemma.

Lemma A.5. *Let $r \in \{1,3,5,\ldots\}$*

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r] \leq 2\mu_{\max}(Q_e)^{\frac{r}{2}}\sqrt{(m+r-1)!}$$

Lemma A.6. *Let $\mathbf{z}(t)$ be any stationary process, and $r \in \mathbb{N}$, then for a stochastic process $\mathbf{s}(t) = \sum_{k=0}^{\infty}\alpha_k\mathbf{z}(t-k)$, with $\sum_{k=0}^{\infty}\|\alpha_k\| \leq +\infty$, the following holds*

$$\mathbf{E}[\|\mathbf{s}(t)\|^r] \leq \left(\sum_{k=0}^{\infty}\|\alpha_k\|\right)^r \mathbf{E}[\|\mathbf{z}(t)\|^r] \tag{26}$$

Proof A.4 (of Lemma A.6).

$$\mathbf{E}[\|\mathbf{s}(t)\|^r] = \mathbf{E}\left[\|\sum_{k=0}^{\infty}\alpha_k\mathbf{z}(t-k)\|^r\right] \le \mathbf{E}\left[\left(\sum_{k=0}^{\infty}\|\alpha_k\|\|\mathbf{z}(t-k)\|\right)^r\right]$$
$$= \mathbf{E}\left[\sum_{k_1=0}^{\infty}\cdots\sum_{k_r=0}^{\infty}\left(\prod_{i=1}^{r}\|\alpha_{k_i}\|\prod_{i=0}^{r}\|\mathbf{z}(t-k_i)\|\right)\right] = \sum_{k_1=0}^{\infty}\cdots\sum_{k_r=0}^{\infty}\left(\prod_{i=1}^{r}\|\alpha_{k_i}\|\mathbf{E}\left[\prod_{i=0}^{r}\|\mathbf{z}(t-k_i)\|\right]\right) \quad (27)$$

By the inequality of arithmetic and geometric means

$$\prod_{i=0}^{r}\|\mathbf{z}(t-k_i)\| \le \frac{1}{r}\sum_{i=1}^{r}\|\mathbf{z}(t-k_i)\|^r \quad (28)$$

then

$$\mathbf{E}\left[\prod_{i=0}^{r}\|\mathbf{z}(t-k_i)\|\right] \le \mathbf{E}\left[\frac{1}{r}\sum_{i=1}^{r}\|\mathbf{z}(t-k_i)\|^r\right] = \frac{1}{r}\sum_{i=1}^{r}\mathbf{E}\left[\|\mathbf{z}(t-k_i)\|^r\right] \quad (29)$$

By assumption $\mathbf{z}(t)$ is stationary, therefore $\mathbf{E}[\|\mathbf{z}(t-k_i)\|^r] = \mathbf{E}[\|\mathbf{z}(t)\|^r]$, i.e. $\mathbf{E}[\|\mathbf{z}(t)\|^r]$ does not depend on k_i, and so we obtain the statement of the lemma

$$\mathbf{E}[\|\mathbf{s}(t)\|^r] \le \mathbf{E}[\|\mathbf{z}(t)\|^r]\sum_{k_1=0}^{\infty}\cdots\sum_{k_r=0}^{\infty}\left(\prod_{i=1}^{r}\|\alpha_{k_i}\|\right) = \left(\sum_{k=0}^{\infty}\|\alpha_k\|\right)^r\mathbf{E}[\|\mathbf{z}(t)\|^r] \quad (30)$$

Lemma A.7. *Let $r \in \mathbb{N}$, then with notation as above the following holds*

$$\mathbf{E}[\|\mathbf{z}_\infty(t) - \mathbf{z}_f(t)\|^r] \le \hat{\gamma}^{rt}\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^r\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \quad (31)$$

Proof A.5 (of Lemma A.7). Notice that the process $\mathbf{s}(t) = \mathbf{z}_\infty(t) - \mathbf{z}_f(t) = \hat{\mathbf{y}}_f(t|0) - \hat{\mathbf{y}}_f(t)$ can be expressed as:

$$\mathbf{s}(t) = \left(\sum_{k=1}^{t}\hat{C}\hat{A}^{k-1}\hat{B}\mathbf{w}(t-k) + \hat{D}\mathbf{w}(t)\right) - \left(\sum_{k=1}^{\infty}\hat{C}\hat{A}^{k-1}\hat{B}\mathbf{w}(t-k) + \hat{D}\mathbf{w}(t)\right) \quad (32)$$
$$= -\sum_{k=t+1}^{\infty}\hat{C}\hat{A}^{k-1}\hat{B}\mathbf{w}(t-k) \quad (33)$$

in the case of $\mathbf{w}(t) = \mathbf{u}(t)$

$$\mathbf{s}(t) = -\sum_{k=t+1}^{\infty}\hat{C}\hat{A}^{k-1}\hat{B}\mathbf{u}(t-k) = \sum_{k=0}^{\infty}\alpha_{k,t}(s,1)\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} \quad (34)$$

with

$$\alpha_{k,t}(s,1) = \begin{cases}\begin{bmatrix}0 & -\hat{C}\hat{A}^{k-1}\hat{B}\end{bmatrix}, & k \ge t+1\\ 0, & k < t+1\end{cases} \quad (35)$$

In the case of $\mathbf{w}(t) = \begin{bmatrix}\mathbf{y}^T(t) & \mathbf{u}^T(t)\end{bmatrix}^T$

$$\mathbf{s}(t) = -\sum_{k=t+1}^{\infty}\hat{C}\hat{A}^{k-1}\hat{B}\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} = \sum_{k=0}^{\infty}\alpha_{k,t}(s,2)\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} \quad (36)$$

with

$$\alpha_{k,t}(s,2) = \begin{cases}-\hat{C}\hat{A}^{k-1}\hat{B}, & k \ge t+1\\ 0, & k < t+1\end{cases} \quad (37)$$

Notice that in both cases we can upper-bound with the same quantity $\|\alpha_{k,t}(s,1)\| \leq \|\alpha_{k,t}(s)\|$, and $\|\alpha_{k,t}(s,2)\| \leq \|\alpha_{k,t}(s)\|$ with

$$\|\alpha_{k,t}(s)\| = \begin{cases} \|\hat{C}\hat{A}^{k-1}\hat{B}\|, & k \geq t+1 \\ 0, & k < t+1 \end{cases} \tag{38}$$

Since $\mathbf{w}(t)$ is a stationary process, and by assumption predictors are stable, i.e. all eigenvalues of $\hat{A}$ are inside unit circle, thus $\sum_{k=0}^{\infty} \|\alpha_{k,t}(s)\| \leq +\infty, \forall t \geq 0$, we apply Lemma A.6, and obtain

$$\mathbf{E}[\|\mathbf{s}(t)\|^r] = \mathbf{E}[\|\mathbf{z}_\infty(t) - \mathbf{z}_f(t)\|^r] \leq \left(\sum_{k=0}^{\infty} \|\alpha_{k,t}(s)\|\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{39}$$

$$\leq \left(\sum_{k=t+1}^{\infty} \|\hat{C}\|\|\hat{A}^{k-1}\|\|\hat{B}\|\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{40}$$

with $\|\hat{A}^k\| \leq \hat{M}\hat{\gamma}^k$, for some $M > 1$ and $\hat{\gamma} \in [\hat{\gamma}^*, 1)$, where $\hat{\gamma}^*$ is the spectral radius of $\hat{A}$, then with a sum of geometric series, we get the statement of the lemma

$$\mathbf{E}[\|\mathbf{z}_\infty(t) - \mathbf{z}_f(t)\|^r] \leq \left(\hat{M}\|\hat{C}\|\|\hat{B}\|\frac{\hat{\gamma}^t}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right]. \tag{41}$$

Lemma A.8. *Let $r \in \mathbb{N}$, then with notation as above the following holds*

$$\mathbf{E}\left[\|\mathbf{z}_\infty(t)\|^r\right] \leq \left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{42}$$

Proof A.6 (of Lemma A.8). Notice that $\mathbf{z}_\infty(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_f(t)$ can be expressed as

In the case of $\mathbf{w}(t) = \mathbf{u}(t)$,

$$\mathbf{z}_\infty(t) = \mathbf{y}(t) - \sum_{k=1}^{\infty} \hat{C}\hat{A}^{k-1}\hat{B}\mathbf{u}(t-k) - \hat{D}\mathbf{u}(t) = \sum_{k=0}^{\infty} \alpha_k(\mathbf{z}_\infty, 1)\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} \tag{43}$$

with

$$\alpha_k(\mathbf{z}_\infty, 1) = \begin{cases} \begin{bmatrix} I & -\hat{D}\end{bmatrix}, & k = 0 \\ \begin{bmatrix} 0 & -\hat{C}\hat{A}^{k-1}\hat{B}\end{bmatrix}, & k > 0 \end{cases} \tag{44}$$

in the case of $\mathbf{w}(t) = [\mathbf{y}^T(t), \mathbf{u}^T(t)]^T$

$$\mathbf{z}_\infty(t) = \mathbf{y}(t) - \sum_{k=1}^{\infty} \hat{C}\hat{A}^{k-1}\hat{B}\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} - \hat{D}\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix} = \sum_{k=0}^{\infty} \alpha_k(\mathbf{z}_\infty, 2)\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} \tag{45}$$

Recall that in this case, we assume $\hat{D} = [0, \hat{D}_{\mathbf{u}}]$, note that $\|\hat{D}\| = \|\hat{D}_{\mathbf{u}}\|$ and thus

$$\alpha_k(\mathbf{z}_\infty, 2) = \begin{cases} \begin{bmatrix} I & -\hat{D}_{\mathbf{u}}\end{bmatrix}, & k = 0 \\ -\hat{C}\hat{A}^{k-1}\hat{B}, & k > 0 \end{cases} \tag{46}$$

Note that in both cases we can upper-bound with the same quantity, i.e. $\|\alpha_k(\mathbf{z}_\infty)\| \leq \|\alpha_k(\mathbf{z}_\infty)\|$, and $\|\alpha_k(\mathbf{z}_\infty, 2)\| \leq \|\alpha_k(\mathbf{z}_\infty)\|$, with

$$\|\alpha_k(\mathbf{z}_\infty)\| \leq \begin{cases} 1 + \|\hat{D}\|, & k = 0 \\ \|\hat{C}\hat{A}^{k-1}\hat{B}\|, & k > 0 \end{cases} \tag{47}$$

Since, in both cases, $\sum_{k=0}^{\infty} \|\alpha_k(\mathbf{z}_\infty)\| \leq +\infty$, due to stability of the predictor, and $\begin{bmatrix}\mathbf{y}^T(t) & \mathbf{u}^T(t)\end{bmatrix}^T$ is stationary, we apply Lemma A.6, to both cases, and upper bound by (47), to obtain an upper-bound for both cases:

$$\mathbf{E}\left[\|\mathbf{z}_\infty(t)\|^r\right] \leq \left(\sum_{k=0}^{\infty} \|\alpha_k(\mathbf{z}_\infty, 1)\|\right)^4 \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{48}$$

$$\leq \left(\|I\| + \|\hat{D}\| + \sum_{k=1}^{\infty} \|\hat{C}\hat{A}^{k-1}\hat{B}\|\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{49}$$

$$\leq \left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{50}$$

Lemma A.9. *Let $r \in \mathbb{N}$, then with notation as above, the following holds*

$$\mathbf{E}\left[\|\mathbf{z}_f(t)\|^r\right] \leq \left(\|I\| + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{51}$$

Proof A.7 (of Lemma A.9). Notice that the process $\mathbf{z}_f(t) = \mathbf{y}(t) - \hat{\mathbf{y}}(t|0)$ can be expressed as:
In the case of $\mathbf{w}(t) = \mathbf{u}(t)$

$$\mathbf{z}_f(t) = \mathbf{y}(t) - \sum_{k=1}^{t} \hat{C}\hat{A}^{k-1}\hat{B}\mathbf{u}(t-k) - \hat{D}\mathbf{u}(t) = \sum_{k=0}^{\infty} \alpha_k(\mathbf{z}_f, 1)\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} \tag{52}$$

with

$$\alpha_k(\mathbf{z}_f, 1) = \begin{cases}\begin{bmatrix} I & -\hat{D}\end{bmatrix}, & k = 0\\ \begin{bmatrix} 0 & -\hat{C}\hat{A}^{k-1}\hat{B}\end{bmatrix}, & 0 < k \leq t\\ 0, & k > t\end{cases} \tag{53}$$

In the case of $\mathbf{w}(t) = [\mathbf{y}^T(t), \mathbf{u}^T(t)]^T$,

$$\mathbf{z}_f(t) = \mathbf{y}(t) - \sum_{k=1}^{t} \hat{C}\hat{A}^{k-1}\hat{B}\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} - \hat{D}\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix} = \sum_{k=0}^{\infty} \alpha_k(\mathbf{z}_f, 2)\begin{bmatrix}\mathbf{y}(t-k)\\ \mathbf{u}(t-k)\end{bmatrix} \tag{54}$$

with

$$\alpha_k(\mathbf{z}_f, 2) = \begin{cases}\begin{bmatrix} I & 0\end{bmatrix} - \hat{D}, & k = 0\\ -\hat{C}\hat{A}^{k-1}\hat{B}, & 0 < k \leq t\\ 0, & k > t\end{cases} \tag{55}$$

Note that for both cases we can upper-bound by the same quantity $\|\alpha_k(\mathbf{z}_f, 1)\| \leq \|\alpha_k(\mathbf{z}_f)\|$, and $\|\alpha_k(\mathbf{z}_f, 2)\| \leq \|\alpha_k(\mathbf{z}_f)\|$, with

$$\|\alpha_k(\mathbf{z}_f)\| = \begin{cases}1 + \|\hat{D}\|, & k = 0\\ \|\hat{C}\hat{A}^{k-1}\hat{B}\|, & 0 < k \leq t\\ 0, & k > t\end{cases} \tag{56}$$

Since by assumption predictors are stable, we apply Lemma A.6 and obtain

$$\mathbf{E}\left[\|\mathbf{z}_f(t)\|^r\right] \leq \left(\sum_{k=0}^{\infty} \|\alpha_k(\mathbf{z}_f)\|\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{57}$$

$$\leq \left(\|I\| + \|\hat{D}\| + \sum_{k=1}^{t} \|\hat{C}\hat{A}^{k-1}\hat{B}\|\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{58}$$

$$\leq \left(\|I\| + \|\hat{D}\| + \hat{M}\|\hat{B}\|\|\hat{C}\| \sum_{k=1}^{t} \hat{\gamma}^{k-1}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{59}$$

$$= \left(\|I\| + \|\hat{D}\| + \hat{M}\|\hat{B}\|\|\hat{C}\| \frac{1-\hat{\gamma}^t}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{60}$$

Notice that $\hat{\gamma}^t > 0, \forall t$, thus we obtain the statement of the lemma

$$\mathbf{E}\left[\|\mathbf{z}_f(t)\|^r\right] \leq \left(\|I\| + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right]. \tag{61}$$

Lemma A.10. *Let $r \in \mathbb{N}$, then with notation as above, the following holds.*

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \leq \|\Sigma_{gen}\|_{\ell_1}^r G_r(\mathbf{e}_g) \tag{62}$$

with

$$\|\Sigma_{gen}\|_{\ell_1} = \|I\| + \sum_{k=1}^{\infty} \|C_g A_g^{k-1} K_g\| \tag{63}$$

$$G_r(\mathbf{e}_g) = \begin{cases} 2^{\frac{r}{2}} \mu_{\max}(Q_e)^{\frac{r}{2}} (n_u + n_y + \frac{r}{2} - 1)!, & r \textit{ is even} \\ 2\mu_{\max}(Q_e)^{\frac{r}{2}} \sqrt{(n_u + n_y + r - 1)!}, & r \textit{ is odd} \end{cases} \tag{64}$$

Proof A.8 (of Lemma A.10). Note that $\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}$ can be expressed as

$$\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix} = \sum_{k=1}^{\infty} C_g A_g^{k-1} K_g \mathbf{e}_g(t-k) + \mathbf{e}_g(t) = \sum_{k=0}^{\infty} \alpha_k(\mathbf{y}, \mathbf{w}) \mathbf{e}_g(t-k) \tag{65}$$

with $\mathbf{e}(t)$ stationary, we apply Lemma A.6 to get

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \leq \left(\sum_{k=0}^{\infty} \|\alpha_k(\mathbf{y}, \mathbf{w})\|\right)^r \mathbf{E}\left[\|\mathbf{e}_g(t)\|^r\right] \tag{66}$$

Let us denote $\|\Sigma_{gen}\|_{\ell_1} = \sum_{k=0}^{\infty} \|\alpha_k(\mathbf{y}, \mathbf{w})\|$, the ℓ_1 norm of the generative system. Furthermore we can apply Lemma A.4 and Lemma A.5 to obtain,

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r] \leq G_r(\mathbf{e}_g) = \begin{cases} 2^{\frac{r}{2}} \mu_{\max}(Q_e)^{\frac{r}{2}} (n_u + n_y + \frac{r}{2} - 1)!, & r \text{ is even} \\ 2\mu_{\max}(Q_e)^{\frac{r}{2}} \sqrt{(n_u + n_y + r - 1)!}, & r \text{ is odd} \end{cases}$$

with this we have the statement of the lemma.

Lemma A.11. *Let $r \in \mathbb{N}$, and $r \geq 0$, then for $a, b \in \mathbb{R}$ the following holds*

$$(a+b)^{2r} \leq 2^{2r-1} a^{2r} + 2^{2r-1} b^{2r} \tag{67}$$

Proof A.9 (of Lemma A.11).

$$(a+b)^{2r} = 2^{2r}\frac{1}{2^{2r}}(a+b)^{2r} = 2^{2r}\left(\frac{1}{2}(a+b)\right)^{2r} \tag{68}$$

since $\phi(x) = x^{2r}$ is convex for $r \geq 0$, we have by definition of convexity

$$\left(\frac{1}{2}(a+b)\right)^{2r} = \phi\left(\frac{a+b}{2}\right) \leq \frac{1}{2}\phi(a) + \frac{1}{2}\phi(b) \tag{69}$$

thus we obtain the statement of the lemma

$$(a+b)^{2r} \leq \frac{2^{2r}}{2}(a^{2r}+b^{2r}) = 2^{2r-1}(a^{2r}+b^{2r}) \tag{70}$$

Lemma A.12. *Let $r \in \mathbb{N}$, then with notation as above, the following holds*

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{(n_u+n_y+r-1)!}{\sqrt{N}}\left(4\bar{G}_{gen}\bar{G}_f(f)\right)^r \tag{71}$$

with

$$\bar{G}_f(f) = \left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{(1-\hat{\gamma})^{\frac{3}{2}}} \tag{72}$$

$$\bar{G}_{gen} = \|\Sigma_{gen}\|_{\ell_1}^2 \mu_{\max}(Q_e) \tag{73}$$

Proof A.10. with $\mathbf{z}_\infty(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_f(t)$, and $\mathbf{z}_f(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_f(t|0)$, we start by applying triangle inequalities

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] = \mathbf{E}\left[\left|\frac{1}{N}\sum_{t=0}^{N-1}\|\mathbf{z}_\infty(t)\|^2 - \|\mathbf{z}_f(t)\|^2\right|^r\right] \leq \mathbf{E}\left[\left(\frac{1}{N}\sum_{t=0}^{N-1}\left|\|\mathbf{z}_\infty(t)\|^2 - \|\mathbf{z}_f(t)\|^2\right|\right)^r\right] \tag{74}$$

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\mathbf{E}\left[\prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t_j)\|^2 - \|\mathbf{z}_f(t_j)\|^2\right|\right] \tag{75}$$

Now using the fact that $|a^2-b^2| = |(a-b)(a+b)| = |a-b|(a+b)$, since $a, b \geq 0$, we get

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\mathbf{E}\left[\prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\|\right|\left(\|\mathbf{z}_\infty(t_j)\| + \|\mathbf{z}_f(t_j)\|\right)\right] \tag{76}$$

We apply Cauchy-Schwarz, i.e. $\mathbf{E}[XY] \leq |\mathbf{E}[XY]| \leq \sqrt{\mathbf{E}[X^2]}\sqrt{\mathbf{E}[Y^2]}$, with $X = \prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\|\right|$, and $Y = \prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\| + \|\mathbf{z}_f(t_j)\|\right)$,

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\mathbf{E}\left[\prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\|\right|^2\right]}\sqrt{\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\| + \|\mathbf{z}_f(t_j)\|\right)^2\right]} \tag{77}$$

For now let's focus on $\mathbf{E}\left[\prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\|\right|^2\right]$, by applying reverse triangle inequality we obtain

$$\mathbf{E}\left[\prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t)\| - \|\mathbf{z}_f(t)\|\right|^2\right] \leq \mathbf{E}\left[\prod_{j=1}^{r}\|\mathbf{z}_\infty(t) - \mathbf{z}_f(t)\|^2\right] \tag{78}$$

now we apply the inequality of arithmetic-geometric means

$$\mathbf{E}\left[\prod_{j=1}^{r}\|\mathbf{z}_\infty(t)-\mathbf{z}_f(t)\|^2\right]\leq\frac{1}{r}\sum_{j=1}^{r}\mathbf{E}[\|\mathbf{z}_\infty(t)-\mathbf{z}_f(t)\|^{2r}] \tag{79}$$

by applying Lemma A.7, we obtain the first term

$$\mathbf{E}\left[\prod_{j=1}^{r}|\|\mathbf{z}_\infty(t_j)\|-\|\mathbf{z}_f(t_j)\||^2\right]\leq\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^{2r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right]\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j} \tag{80}$$

Now for the second term $\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^2\right]$, we apply the inequality of arithmetic-geometric means

$$\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^2\right]\leq\frac{1}{r}\sum_{j=1}^{r}\mathbf{E}\left[\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^{2r}\right] \tag{81}$$

By Lemma A.11, we obtain

$$\frac{1}{r}\sum_{j=1}^{r}\mathbf{E}\left[\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^{2r}\right]\leq\frac{2^{2r-1}}{r}\sum_{j=1}^{r}\left(\mathbf{E}\left[\|\mathbf{z}_\infty(t_j)\|^{2r}\right]+\mathbf{E}\left[\|\mathbf{z}_f(t_j)\|^{2r}\right]\right) \tag{82}$$

By Lemma A.8 and Lemma A.9, we obtain

$$\frac{2^{2r-1}}{r}\sum_{j=1}^{r}\left(\mathbf{E}\left[\|\mathbf{z}_\infty(t_j)\|^{2r}\right]+\mathbf{E}\left[\|\mathbf{z}_f(t_j)\|^{2r}\right]\right)\leq\frac{2^{2r}}{r}\sum_{j=1}^{r}\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^{2r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \tag{83}$$

$$=2^{2r}\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^{2r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \tag{84}$$

Now taking (84) and (80) back to (77), we have

$$\begin{aligned}\mathbf{E}[\|V_N(f)-\hat{\mathcal{L}}_N(f)\|^r]&\leq\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\mathbf{E}\left[\prod_{j=1}^{r}|\|\mathbf{z}_\infty(t_j)\|-\|\mathbf{z}_f(t_j)\||^2\right]}\sqrt{\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^2\right]}\\&\leq\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^{2r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right]\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j}}\\&\quad\cdot\sqrt{2^{2r}\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^{2r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right]}\end{aligned} \tag{85}$$

$$\begin{aligned}\mathbf{E}[\|V_N(f)-\hat{\mathcal{L}}_N(f)\|^r]&\leq 2^r\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^{r}\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^{r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right]\\&\quad\cdot\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j}}\end{aligned} \tag{86}$$

Note that we can write

$$\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j}}=\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\phi(\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j}) \tag{87}$$

thus we can apply Jensen's inequality for concave function $\phi(x) = \sqrt{x}$, i.e. $\phi\left(\frac{1}{\|S\|}\sum_{i\in S} x_i\right) \geq \frac{1}{\|S\|}\sum_{i\in S}\phi(x_i)$, thus we obtain

$$\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j}} \leq \sqrt{\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j}} \tag{88}$$

Now by commuting the sums we get

$$\sqrt{\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\frac{1}{r}\sum_{j=1}^{r}\hat{\gamma}^{2rt_j}} = \sqrt{\frac{1}{r}\sum_{j=1}^{r}\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\hat{\gamma}^{2rt_j}} \tag{89}$$

now notice that $\hat{\gamma}^{2rt_j}$ only depend on one sum, for which we can use the sum of geometric series, after which the same term will be repeated N^{r-1} times, therefore

$$\sqrt{\frac{1}{r}\sum_{j=1}^{r}\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\hat{\gamma}^{2rt_j}} = \sqrt{\frac{1}{r}\sum_{j=1}^{r}\frac{N^{r-1}}{N^r}\frac{1-\hat{\gamma}^{2rN}}{1-\hat{\gamma}^{2r}}} = \frac{1}{\sqrt{N}}\sqrt{\frac{1-\hat{\gamma}^{2rN}}{1-\hat{\gamma}^{2r}}} \tag{90}$$

since $\hat{\gamma}^{2rN} \geq 0$, and $(1-\hat{\gamma})^{\frac{r}{2}} \leq (1-\hat{\gamma}^{2r})^{\frac{1}{2}}$, since

$$(1-\hat{\gamma})^{\frac{r}{2}} \leq ((1-\hat{\gamma}^r)(1+\hat{\gamma}^r))^{\frac{1}{2}} \tag{91}$$

$$1 \leq (1+\hat{\gamma}^r) \tag{92}$$

we obtain

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{2^r}{\sqrt{N}}\left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{(1-\hat{\gamma})^{\frac{3}{2}}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \tag{93}$$

We can apply Lemma A.10, to get

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \leq \|\Sigma_{gen}\|_{\ell_1}^{2r} G_{2r}(\mathbf{e}_g) \tag{94}$$

since $2r$ is always even, then

$$G_{2r}(\mathbf{e}_g) = 2^r \mu_{\max}(Q_e)^r (n_u + n_y + r - 1)! \tag{95}$$

and with this we obtain the statement of the lemma

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{2^{2r}}{\sqrt{N}}\left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{(1-\hat{\gamma})^{\frac{3}{2}}}\right)^r \cdot \|\Sigma_{gen}\|_{\ell_1}^{2r}\mu_{\max}(Q_e)^r(n_u + n_y + r - 1)! \tag{96}$$

with some algebraic manipulation we get

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{(n_u + n_y + r - 1)!}{\sqrt{N}}\left(4\left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{(1-\hat{\gamma})^{\frac{3}{2}}}\|\Sigma_{gen}\|_{\ell_1}^2\mu_{\max}(Q_e)\right)^r \tag{97}$$

Lemma A.13. *With notation as above for* $0 < \lambda < \frac{1}{4n_w\bar{G}_{gen}\bar{G}_f(f)}$ *following holds*

$$\mathbf{E}[e^{\lambda|V_N(f) - \hat{\mathcal{L}}_N(f)|}] \leq 1 + \frac{(n_y + n_u)!}{\sqrt{N}}\frac{4\lambda\bar{G}_{gen}\bar{G}_f(f)}{1 - 4\lambda(n_y + n_u)\bar{G}_{gen}\bar{G}_f(f)} \tag{98}$$

Proof A.11 (of Lemma A.13). with $X = \lambda|V_N(f) - \hat{\mathcal{L}}_N(f)|$

$$\mathbf{E}[e^{\lambda(V_N(f)-\hat{\mathcal{L}}_N(f))}] = 1 + \sum_{r=1}^{\infty} \frac{\lambda^r}{r!}\mathbf{E}[|V_N(f) - \hat{\mathcal{L}}_N(f)|^r] \leq 1 + \sum_{r=1}^{\infty} \frac{\lambda^r}{r!} \frac{(n_u + n_y + r - 1)!}{\sqrt{N}} \left(4\bar{G}_{gen}\bar{G}_f(f)\right)^r \tag{99}$$

Furthermore, with $n_w = n_u + n_y$

$$\frac{(n_w + r - 1)!}{r!} = n_w! \frac{n_w + 1}{2} \frac{n_w + 2}{3} \cdots \frac{n_w + r - 1}{r}$$

and as $\frac{n_w+r-1}{r} \leq n_w$, for all $r \geq 1$, then

$$\frac{(n_w + r - 1)!}{r!} \leq n_w! \, (n_w)^{r-1} = n_w! \frac{(n_w)^r}{n_w} = \frac{n_w!}{n_w} (n_w)^r = (n_w - 1)!(n_w)^r.$$

this allows us to write

$$\mathbf{E}[e^{\lambda(V_N(f)-\hat{\mathcal{L}}_N(f))}] \leq 1 + \frac{(n_w - 1)!}{\sqrt{N}} \sum_{r=1}^{\infty} \left(4\lambda n_w \bar{G}_{gen}\bar{G}_f(f)\right)^r \tag{100}$$

the infinite sum is absolutely convergent if

$$4\lambda n_w \bar{G}_{gen}\bar{G}_f(f) < 1$$

that means that

$$0 < \lambda < \frac{1}{4n_w \bar{G}_{gen}\bar{G}_f(f)} \tag{101}$$

under this condition we can write

$$\mathbf{E}[e^{\lambda(V_N(f)-\hat{\mathcal{L}}_N(f))}] \leq 1 + \frac{(n_w - 1)!}{\sqrt{N}} \frac{4\lambda n_w \bar{G}_{gen}\bar{G}_f(f)}{1 - 4\lambda n_w \bar{G}_{gen}\bar{G}_f(f)} = 1 + \frac{n_w!}{\sqrt{N}} \frac{4\lambda \bar{G}_{gen}\bar{G}_f(f)}{1 - 4\lambda n_w \bar{G}_{gen}\bar{G}_f(f)} \tag{102}$$

Lemma A.14. *Let $\mathbf{y}_\nu(t), \hat{\mathbf{y}}_{f,\nu}(t), \hat{\mathbf{y}}_{f,\nu}(t|s) \in \mathbb{R}^1$ denote the ν'th component of $\mathbf{y}(t), \hat{\mathbf{y}}_f(t), \hat{\mathbf{y}}_f(t|s)$ respectively,*

$$\mathcal{L}_\nu(f) \triangleq \mathbf{E}[(\hat{\mathbf{y}}_{f,\nu}(t) - \mathbf{y}_\nu(t))^2] = \lim_{s \to -\infty} \mathbf{E}[(\hat{\mathbf{y}}_{f,\nu}(t|s) - \mathbf{y}_\nu(t))^2] \tag{103}$$

$$V_{N,\nu}(f) \triangleq \frac{1}{N} \sum_{t=0}^{N-1} (\hat{\mathbf{y}}_{f,\nu}(t) - \mathbf{y}_\nu(t))^2 \tag{104}$$

and let $\sigma(r)$, be such that the following holds.

$$\sigma(r) \geq \sup_{t,k,l} \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \tag{105}$$

$$\mathbf{e}(t,k,j) = \begin{cases} Q_e - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k = j \\ -\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k \neq j \end{cases} \tag{106}$$

Then the raw moments are bounded

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N}\sigma(r)4(r-1)G_e(f)^{2r} \tag{107}$$

Proof A.12 (Proof of Lemma A.14). The prediction error can be expressed as

$$(\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t)) = \sum_{k=0}^{\infty} \alpha_k \mathbf{e}_g(t-k)$$

with

$$\alpha_k = \alpha_k(\nu) = \begin{cases} D_{e_\nu}, & k = 0 \\ C_{e_\nu} A_e^{k-1} K_e, & k > 0 \end{cases}$$

where $D_{e_\nu} = \mathbf{1}_\nu D_e$, and $C_{e_\nu} = \mathbf{1}_\nu C_e$ denote the ν'th row of matrices D_e, C_e respectively. Then generalised loss $\mathcal{L}_\nu(f)$ for component ν is expressed as

$$\begin{aligned} \mathcal{L}_\nu(f) &= \mathbf{E}[(\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t))^2] \\ &= \mathbf{E}\left[\text{trace}\left(\left(\sum_{k=0}^{\infty} \alpha_k \mathbf{e}_g(t-k)\right)\left(\sum_{k=0}^{\infty} \alpha_k \mathbf{e}_g(t-k)\right)^T\right)\right] \\ &= \sum_{k=0}^{\infty} \alpha_k Q_e \alpha_k^T \end{aligned}$$

and infinite horizon prediction loss is

$$\begin{aligned} V_{N,\nu}(f) &= \frac{1}{N}\sum_{t=0}^{N-1} (\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t))^2 \\ \mathcal{L}_\nu(f) - V_{N,\nu}(f) &= \frac{1}{N}\sum_{t=0}^{N-1}\left(\sum_{k=0}^{\infty} \alpha_k Q_e \alpha_k^T - \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \alpha_k \mathbf{e}_g(t-k)\mathbf{e}_g(t-j)\alpha_k^T\right) \\ &= \frac{1}{N}\sum_{t=0}^{N-1}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \alpha_k \mathbf{e}(t,k,j)\alpha_j^T \\ \mathbf{e}(t,k,j) &= \begin{cases} \text{trace}(Q_e) - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k = j \\ -\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k \neq j \end{cases} \end{aligned}$$

For ease of notation let us define

$$\mathbf{z}(t,k,j) = \alpha_k \mathbf{e}(t,k,j)\alpha_j^T$$

then

$$\begin{aligned} &\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \\ &= \frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sum_{k_1,j_1=0}^{\infty}\cdots\sum_{k_r,j_r=0}^{\infty} \mathbf{E}\left[\prod_{l=1}^{r} z(t_l,k_l,j_l)\right] \end{aligned}$$

Note that, with i.i.d. innovation noise $\mathbf{e}_g(t)$, if

$$\begin{aligned} &t_r - k_r \notin \{t_i - k_i, t_i - j_i\}_{i=1}^{r-1} \\ &\quad \wedge\, t_r - j_r \notin \{t_i - k_i, t_i - j_i\}_{i=1}^{r-1} \end{aligned}$$

or similarly

$$\{t_r - k_r, t_r - j_r\} \cap \{t_i - k_i, t_i - j_i\}_{i=1}^{r-1} = \emptyset \tag{108}$$

then $\mathbf{z}(t_r,k_r,j_r)$ is independent of $\mathbf{z}(t_i,k_i,j_i)$. Moreover, notice that $E(\mathbf{z}(t_r,k_r,j_r)] = 0$. Hence, if (108), it holds that

$$\mathbf{E}\left[\prod_{l=1}^{r} z(t_l,k_l,j_l)\right] = \mathbf{E}\left[\prod_{l=1}^{r-1} \mathbf{z}(t_l,k_l,j_l)\right] \underbrace{\mathbf{E}[\mathbf{z}(t_r,k_r,j_r)]}_{=0} = 0. \tag{109}$$

Let us denote

$$\mathcal{Z} = \{t_i - k_i + k_r, t_i - j_i + k_r, t_i - k_i + j_r, t_i - j_i + j_r\}_{i=1}^{r-1}.$$

Then using (109) for those $\{t_l, k_l, j_l\}_{l=1}^r$ which satisfy (108), it follows that

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] = \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_{r-1}=0}^{N-1} \sum_{k_1,j_1=0}^{\infty} \cdots \sum_{k_r,j_r=0}^{\infty} \sum_{t_r \in \mathcal{Z}} \mathbf{E}\left[\prod_{l=1}^r z(t_l, k_l, j_l)\right]. \tag{110}$$

Note that

$$\mathbf{E}\left[\prod_{l=1}^r z(t_l, k_l, j_l)\right] \leq \left|\mathbf{E}\left[\prod_{l=1}^r z(t_l, k_l, j_l)\right]\right| \leq \mathbf{E}\left[\prod_{l=1}^r |z(t_l, k_l, j_l)|\right].$$

Let us focus on $|\mathbf{z}(t_i, k_i, j_i)|$:

$$|\mathbf{z}(t_l, k_l, j_l)| \leq \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \|\mathbf{e}(t_l, k_l, j_l)\|_2$$
$$\mathbf{E}\left[\prod_{l=1}^r |\mathbf{z}(t_l, k_l, j_l)|\right] \leq \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \mathbf{E}\left[\prod_{l=1}^r \|\mathbf{e}(t_l, k_l, j_l)\|_2\right]$$

Then using Arithmetic Mean-Geometric Mean Inequality, (Steele, 2004) we have

$$\mathbf{E}\left[\prod_{l=1}^r \|\mathbf{e}(t_l, k_l, j_l)\|\right] \leq \frac{1}{r} \sum_{l=1}^r \mathbf{E}[\|\mathbf{e}(t_l, k_l, j_l)\|_2^r] \tag{111}$$

Now, let $\sigma(r)$, be such that the following holds.

$$\sigma(r) \geq \sup_{t,k,l} \mathbf{E}[\|\mathbf{e}(t, k, l)\|_2^r] \tag{112}$$

Then, $\frac{1}{r} \sum_{l=1}^r \mathbf{E}[\|\mathbf{e}(t_l, k_l, j_l)\|_2^r] \leq \sigma(r)$ and then from (111) it follows that

$$\mathbf{E}\left[\prod_{l=1}^r |\mathbf{e}(t_l, k_l, j_l)|\right] \leq \sigma(r) \tag{113}$$

Combining this with (110), it follows that

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_{r-1}=0}^{N-1} \sum_{k_1,j_1=0}^{\infty} \cdots \sum_{k_r,j_r=0}^{\infty} \sum_{t_r \in \mathcal{Z}} \sigma(r) \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \tag{114}$$

and the quantity $\sigma(r) \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2$ does not depend on t_r. Moreover

$$\sum_{t_r \in \mathcal{Z}} \sigma(r) \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \leq \sigma(r) \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 |\mathcal{Z}|,$$

where $|\mathcal{Z}|$ is the cardinality of the set $\mathcal{Z}$. Note $|\mathcal{Z}| \leq 4(r-1)$, therefore

$$\sum_{t_r \in \mathcal{Z}} \sigma(r) \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \leq \sigma(r) \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 4(r-1),$$

Combining the latter inequality with (114), it follows that

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_{r-1}=0}^{N-1} \sigma(r) 4(r-1) \sum_{k_1,j_1=0}^{\infty} \cdots \sum_{k_r,j_r=0}^{\infty} \prod_{l=1}^r \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \tag{115}$$

Now notice

$$G_{e,\nu}(f)^{2r} = \left(\sum_{k=0}^{\infty} \|\alpha_k\|_2\right)^{2r} = \left(\sum_{k,j=0}^{\infty} \|\alpha_k\|_2 \|\alpha_j\|_2\right)^{r}$$
$$= \sum_{k_1,j_1=0}^{\infty} \cdots \sum_{k_r,j_r=0}^{\infty} \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2$$

therefore we obtain

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_{r-1}=0}^{N-1} \sigma(r) 4(r-1) G_{e,\nu}(f)^{2r}$$
$$\leq \frac{1}{N^r} N^{r-1} \sigma(r) 4(r-1) G_{e,\nu}(f)^{2r}$$
$$\leq \frac{1}{N} \sigma(r) 4(r-1) G_{e,\nu}(f)^{2r}$$

and since

$$\|\alpha_k(\nu)\| = \begin{cases} \|\mathbf{1}_\nu D_e\| \leq \|D_e\|, & k = 0 \\ \|\mathbf{1}_\nu C_e A_e^{k-1} K_e\| \leq \|C_e A_e^{k-1} K_e\|, & k > 0 \end{cases}$$

then

$$G_{e,\nu} \leq G_e = \|D_e\| + \sum_{k=1}^{\infty} \|C_e A_e^{k-1} K_e\| \tag{116}$$

and since $2r > 1$ we obtain the statement of the lemma

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N} \sigma(r) 4(r-1) G_e(f)^{2r} \tag{117}$$

Lemma A.15. *with notation as above the following holds*

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \leq \frac{n_y^r}{N} \sigma(r) 4(r-1) G_e(f)^{2r} \tag{118}$$

Proof A.13 (of Lemma A.15). By definition

$$\mathcal{L}(f) = \mathbf{E}[(\mathbf{y}(t) - \hat{\mathbf{y}}_f(t))^T (\mathbf{y}(t) - \hat{\mathbf{y}}_f(t))] = \sum_{\nu=1}^{n_y} \mathbf{E}[(\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t))^2] = \sum_{\nu=1}^{n_y} \mathcal{L}_\nu(f) \tag{119}$$

$$V_N(f) = \frac{1}{N} \sum_{t=0}^{N-1} (\mathbf{y}(t) - \hat{\mathbf{y}}_f(t))^T (\mathbf{y}(t) - \hat{\mathbf{y}}_f(t)) = \sum_{\nu=1}^{n_y} \frac{1}{N} \sum_{t=0}^{N-1} (\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t))^2 = \sum_{\nu=1}^{n_y} V_{N,\nu}(f) \tag{120}$$

$$\tag{121}$$

then

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] = \mathbf{E}\left[\left(\sum_{\nu=1}^{n_y} \mathcal{L}_\nu(f) - V_{N,\nu}(f)\right)^r\right] = \sum_{\nu_1}^{n_y} \cdots \sum_{\nu_r}^{n_y} \mathbf{E}\left[\prod_{i=1}^{r} (\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f))\right] \tag{122}$$

Then using Arithmetic Mean-Geometric Mean Inequality, (Steele, 2004), we get $\prod_{i=1}^{r} (\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f)) \leq \frac{1}{r} \sum_{i=1}^{r} (\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f))^r$, and thus

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \leq \sum_{\nu_1=1}^{n_y} \cdots \sum_{\nu_r=1}^{n_y} \frac{1}{r} \sum_{i=1}^{r} \mathbf{E}\left[(\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f))^r\right] \tag{123}$$

From Lemma A.14, we have $\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N}\sigma(r)4(r-1)G_e(f)^{2r}$, thus

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \leq \sum_{\nu_1=1}^{n_y} \cdots \sum_{\nu_r=1}^{n_y} \frac{1}{r}\sum_{i=1}^{r} \frac{1}{N}\sigma(r)4(r-1)G_e(f)^{2r} \tag{124}$$

$$= \frac{n_y^r}{N}\sigma(r)4(r-1)G_e(f)^{2r} \tag{125}$$

Lemma A.16. *let $m = n_u + n_y$, then for $r \geq 2$, the quantity*

$$\sigma(r) = \max\left\{(\mu_{\max}(Q_e)^r 4(m+r-1)!), (\mu_{\max}(Q_e)^r 3^r (m+r-1)!)\right\} = \mu_{\max}(Q_e)^r 3^r (m+r-1)!$$

satisfies

$$\sigma(r) \geq \sup_{t,k,l} \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]$$

Proof A.14 (Proof of Lemma A.16). Recall that

$$\mathbf{e}(t,k,j) = \begin{cases} Q_e - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k = j \\ -\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k \neq j \end{cases}$$

First let us take the case when $k \neq j$. Then

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] = \mathbf{E}[\| - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j)\|_2^r]$$

Again as $\mathbf{e}_g(t)$ is i.i.d. we have

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \mathbf{E}[\|\mathbf{e}_g(t-k)\|_2^r]\mathbf{E}[\|\mathbf{e}_g(t-j)\|_2^r]$$

and due to stationarity of $\mathbf{e}_g(t)$, we have $\mathbf{E}[\|\mathbf{e}_g(t-k)\|_2^r] = \mathbf{E}[\|\mathbf{e}_g(t-j)\|_2^r]$, therefore

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]^2$$

and again due to stationarity of $\mathbf{e}_g(t)$, the moments do not depend on t, and using Lemma A.5 we obtain

$$\sigma(r) \geq \mu_{\max}(Q_e)^r 4((m+r-1)!) \geq \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]^2$$

Now let us take the case when $k = j$. Then

$$\begin{aligned}
\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] &= \mathbf{E}[\|Q_e - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-k)\|_2^r] \\
&\leq \mathbf{E}[(\|Q_e\|_2 + \|\mathbf{e}_g(t)\|_2^2)^r] \\
&= \mathbf{E}\left[\sum_{j=0}^{r}\binom{r}{j}\|Q_e\|_2^{r-j}\|\mathbf{e}_g(t)\|_2^{2j}\right] \\
&= \sum_{j=0}^{r}\binom{r}{j}\|Q_e\|_2^{r-j}\mathbf{E}\|\mathbf{e}_g(t)\|_2^{2j}]
\end{aligned}$$

As Q_e is a positive definite matrix,$\|Q_e\|_2 = \mu_{max}(Q_e)$, and hence

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \sum_{j=0}^{r}\binom{r}{j}\mu_{\max}(Q_e)^{r-j}\mathbf{E}\|\mathbf{e}_g(t)\|_2^{2j}]$$

using Lemma A.4 we obtain

$$\begin{aligned}
\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] &\leq \sum_{j=0}^{r}\binom{r}{j}\mu_{\max}(Q_e)^{r-j}\mu_{\max}(Q_e)^j 2^j(m+j-1)! \\
&\leq \mu_{\max}(Q_e)^r \sum_{j=0}^{r}\binom{r}{j}2^j(m+j-1)!.
\end{aligned}$$

Since for $j \leq r$, $(m+j-1)! \leq (m+r-1)!$, hence

$$\mathbf{E}\|\mathbf{e}(t,k,l)\|_2^{2r}] \leq \mu_{\max}(Q_e)^r (m+r-1)! \sum_{j=0}^{r} \binom{r}{j} 2^j$$

Notice $3^r = (1+2)^r = \sum_{j=0}^{r} \binom{r}{j} 2^j$, hence

$$\mathbf{E}\|\mathbf{e}_g(t,k,l)\|_2^{2r}] \leq \mu_{\max}(Q_e)^r 3^r (m+r-1)!$$

Hence,

$$\begin{aligned}\sigma(r) = \max\{&\mu_{\max}(Q_e)^r 4(m+r-1)!, \\ &\mu_{\max}(Q_e)^r 3^r (m+r-1)!\}.\end{aligned}$$

As we are interested in moments higher or equal to two, i.e. $r \geq 2$, then

$$\sigma(r) = \mu_{\max}(Q_e)^r 3^r (m+r-1)!.$$

Lemma A.17. *For $\lambda \leq \left(3(m+1)n_y \mu_{\max}(Q_e) G_e(f)^2\right)^{-1}$, the moment generating function is bounded*

$$\mathbf{E}\left[e^{\lambda(\mathcal{L}(f)-V_N(f))}\right] \leq 1 + \frac{2}{N} \frac{(m+1)! \left(3\lambda n_y \mu_{\max}(Q_e) G_e(f)^2\right)^2}{(1-3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2)} \tag{126}$$

Proof A.15 (Proof of Lemma A.17). We can bound the moment generating function via series expansion. First note that $\mathbf{E}[\mathcal{L}(f) - V_N(f)] = 0$, and hence

$$\mathbf{E}\left[e^{\lambda(\mathcal{L}(f)-V_N(f))}\right] = 1 + \lambda \mathbf{E}[\mathcal{L}(f) - V_N(f)] + \sum_{r=2}^{\infty} \frac{\lambda^r}{r!} E[(\mathcal{L}(f) - V_N(f))^r].$$

Then using Lemma A.15 we get

$$\mathbf{E}\left[e^{\lambda(\mathcal{L}(f)-V_N(f))}\right] \leq 1 + \sum_{r=2}^{\infty} \frac{\lambda^r}{r!} \frac{n_y^r}{N} \sigma(r) 4(r-1) G_e(f)^{2r} \tag{127}$$

Now using Lemma A.16 we obtain

$$\mathbf{E}\left[e^{\lambda(\mathcal{L}(f)-V_N(f))}\right] \leq 1 + \frac{1}{N} \sum_{r=2}^{\infty} \frac{(m+r-1)!}{r!} 4(r-1) \left(3n_y \lambda \mu_{\max}(Q_e) G_e(f)^2\right)^r$$

Notice that $4(r-1) \leq 2^r$, for $r \in \mathbb{N}$. Furthermore

$$\frac{(m+r-1)!}{r!} = m! \frac{m+1}{2} \frac{m+2}{3} \cdots \frac{m+r-1}{r}$$

and as $\frac{m+r-1}{r} \leq \frac{m+1}{2}$, for all $r \geq 2$, then

$$\frac{(m+r-1)!}{r!} \leq m! \left(\frac{m+1}{2}\right)^{r-1} = m! \frac{\left(\frac{m+1}{2}\right)^r}{\frac{m+1}{2}} = 2\frac{m!}{m+1} \left(\frac{m+1}{2}\right)^r.$$

Hence, we can derive the following inequality:

$$\mathbf{E}\left[e^{\lambda(\mathcal{L}(f)-V_N(f))}\right] \leq 1 + \frac{2}{N} \frac{m!}{m+1} \sum_{r=2}^{\infty} \left(3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2\right)^r.$$

Notice that if

$$|3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2| < 1,$$

then the infinite sum $\sum_{r=2}^{\infty} \left(3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2\right)^r$ is absolutely convergent, and

$$\sum_{r=2}^{\infty} \left(3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2\right)^r = \frac{\left(3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2\right)^2}{1 - 3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2}$$

To sum up, if

$$\lambda \leq \left(3(m+1) n_y \mu_{\max}(Q_e) G_e(f)^2\right)^{-1}.$$

then

$$\begin{aligned}
\mathbf{E}\left[e^{\lambda(\mathcal{L}(f) - V_N(f))}\right] &\leq 1 + \frac{2}{N} \frac{m!}{m+1} \frac{\left(3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2\right)^2}{1 - 3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2} \\
&\leq 1 + \frac{2}{N} \frac{(m+1)! \left(3\lambda n_y \mu_{\max}(Q_e) G_e(f)^2\right)^2}{(1 - 3(m+1)\lambda n_y \mu_{\max}(Q_e) G_e(f)^2)}.
\end{aligned}$$

Lemma A.18. *For measurable functions $X(f), Y(f)$ on $\mathcal{F}$, With probability at least $1-\delta$, the following holds*

$$\forall \rho: \quad E_{f\sim\hat{\rho}} X(f) \leq \; E_{f\sim\hat{\rho}} Y(f) + \frac{1}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \Psi_\pi(\lambda, N)\right], \tag{128}$$

with

$$\Psi_\pi(\lambda, N) = \ln E_{f\sim\pi} \mathbf{E}[e^{\lambda(X(f) - Y(f))}] \tag{129}$$

Proof A.16 (of Lemma A.18). Let us apply the Donsker & Varadhan variational formula to the function $\lambda(X(f) - Y(f))$ it then follows that

$$\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}} X(f) - \lambda E_{f\sim\hat{\rho}} Y(f) - KL(\hat{\rho}\|\pi)) = \ln E_{f\sim\pi} e^{\lambda(X(f) - Y(f))}, \tag{130}$$

In particular,

$$e^{\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}} X(f) - \lambda E_{f\sim\hat{\rho}} Y(f) - KL(\hat{\rho}\|\pi))} = e^{\ln E_{f\sim\pi} e^{\lambda(X(f) - Y(f))}} = E_{f\sim\pi} e^{\lambda(X(f) - Y(f))} \tag{131}$$

and hence

$$\begin{aligned}
\mathbf{E}[e^{\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}} X(f) - \lambda E_{f\sim\hat{\rho}} Y(f) - KL(\hat{\rho}\|\pi))}] = \mathbf{E}[E_{f\sim\pi} e^{\lambda(X(f) - Y(f))}] = \\
E_{f\sim\pi} \mathbf{E}[e^{\lambda(X(f) - Y(f))}] = e^{\Psi_\pi(\lambda, N)}
\end{aligned} \tag{132}$$

with

$$\Psi_\pi(\lambda, N) = \ln E_{f\sim\pi} \mathbf{E}[e^{\lambda(X(f) - Y(f))}] \tag{133}$$

Hence,

$$\mathbf{E}[e^{\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}} X(f) - \lambda E_{f\sim\hat{\rho}} Y(f) - KL(\hat{\rho}\|\pi)}] e^{-\Psi_\pi(\lambda, N)} = 1 \tag{134}$$

Since

$$\begin{aligned}
\mathbf{E}[e^{\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}} X(f) - \lambda E_{f\sim\hat{\rho}} Y(f) - KL(\hat{\rho}\|\pi)}] e^{-\Psi_\pi(\lambda, N)} = \\
\mathbf{E}[e^{\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}} X(f) - \lambda E_{f\sim\hat{\rho}} Y(f) - KL(\hat{\rho}\|\pi) - \Psi_\pi(\lambda, N)}]
\end{aligned} \tag{135}$$

it follows that

$$\mathbf{E}[e^{\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}} X(f) - \lambda E_{f\sim\hat{\rho}} Y(f) - KL(\hat{\rho}\|\pi)) - \Psi_\pi(\lambda, N)}] = 1 \tag{136}$$

By Chernoff's bound applied to the random variable $\mathcal{X} = \sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}}(f) - \lambda E_{f\sim\hat{\rho}}Y(f) - KL(\hat{\rho}\|\pi)) - \Psi_\pi(\lambda, N)$ it then follows that for any $a > 0$

$$\mathbf{P}(\mathcal{X} \geq a) \leq \frac{E[e^{\mathcal{X}}]}{e^a} \leq e^{-a}$$

By choosing $a = \ln \frac{1}{\delta}$, it follows that

$$\mathbf{P}(\mathcal{X} \geq \ln \frac{1}{\delta}) \leq \delta$$

and hence,

$$\mathbf{P}(\mathcal{X} \leq \ln \frac{1}{\delta}) \geq 1 - \delta$$

By substituting the definition of $\mathcal{X}$ and regrouping the terms, it then follows that

$$\mathbf{P}(\sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}}X(f) - \lambda E_{f\sim\hat{\rho}}Y(f) - KL(\hat{\rho}\|\pi)) \leq \ln \frac{1}{\delta} + \Psi_\pi(\lambda, N)) \geq 1 - \delta$$

Note that

$$\{\omega \mid \sup_{\hat{\rho}}(\lambda E_{f\sim\hat{\rho}}X(f) - \lambda E_{f\sim\hat{\rho}}Y(f)(\omega) - KL(\hat{\rho}\|\pi)) \leq \ln \frac{1}{\delta} + \Psi_\pi(\lambda, N)\} =$$
$$\{\omega \mid \forall \hat{\rho} : E_{f\sim\hat{\rho}}X(f) \leq E_{f\sim\hat{\rho}}Y(f)(\omega) + \frac{1}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln \frac{1}{\delta} + \Psi_\pi(\lambda, N)\right\}$$

and hence it then follows that with probability at least $1 - \delta$, the following holds

$$\forall \rho : \quad E_{f\sim\hat{\rho}}X(f) \leq \; E_{f\sim\hat{\rho}}Y(f) + \frac{1}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln \frac{1}{\delta} + \Psi_\pi(\lambda, N)\right], \tag{137}$$

Corollary A.19. *By Lemma A.18, and Lemma A.17, for* $0 < \lambda \leq \inf_{f\in\mathcal{F}} \left(3(m+1)n_y\mu_{\max}(Q_e)G_e(f)^2\right)^{-1}$*, with* $\mathcal{M}_\pi$*, denoting the set of all absolutely continuous probability densities w.r.t.* π*, then with probability at least* $1 - \delta$*, the following holds*

$$\forall \rho \in \mathcal{M}_\pi : \quad E_{f\sim\hat{\rho}}\mathcal{L}(f) \leq E_{f\sim\hat{\rho}}V_N(f) + \frac{1}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln \frac{1}{\delta} + \widehat{\Psi}_{\pi,1}(\lambda, N)\right], \tag{138}$$

with

$$\widehat{\Psi}_{\pi,1}(\lambda, N) \triangleq \ln E_{f\sim\pi}\left(1 + \frac{2}{N}\frac{(m+1)!\left(3\lambda n_y\mu_{\max}(Q_e)G_e(f)^2\right)^2}{(1 - 3(m+1)\lambda n_y\mu_{\max}(Q_e)G_e(f)^2)}\right) \tag{139}$$

Corollary A.20. *By Lemma A.18, and Lemma A.13, for* $0 < \lambda \leq \inf_{f\in\mathcal{F}} \left(4n_w\bar{G}_{gen}\bar{G}_f(f)\right)^{-1}$*, with* $\mathcal{M}_\pi$*, denoting the set of all absolutely continuous probability densities w.r.t.* π*, then with probability at least* $1 - \delta$*, the following holds*

$$\forall \rho \in \mathcal{M}_\pi : \quad E_{f\sim\hat{\rho}}V_N(f) \leq E_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + \frac{1}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln \frac{1}{\delta} + \widehat{\Psi}_{\pi,2}(\lambda, N)\right], \tag{140}$$

with

$$\widehat{\Psi}_{\pi,2}(\lambda, N) \triangleq \ln E_{f\sim\pi}\left(1 + \frac{(n_y + n_u)!}{\sqrt{N}}\frac{4\lambda\bar{G}_{gen}\bar{G}_f(f)}{1 - 4\lambda(n_y + n_u)\bar{G}_{gen}\bar{G}_f(f)}\right) \tag{141}$$

Lemma A.21. *For*

$$0 < \tilde{\lambda} \leq \frac{1}{2}\Big(\sup_{f\in\mathcal{F}} \max\{3(m+1)n_y\mu_{\max}(Q_e)G_e(f)^2, 4n_w\bar{G}_{gen}\bar{G}_f(f)\}\Big)^{-1} \tag{142}$$

with probability at least $1-2\delta$*, the following holds*

$$\forall \rho \in \mathcal{M}_\pi : \quad E_{f\sim\hat{\rho}}\mathcal{L}(f) \leq E_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + \frac{1}{\tilde{\lambda}}\left[KL(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \frac{\widehat{\Psi}_{\pi,2}(2\tilde{\lambda},N) + \widehat{\Psi}_{\pi,1}(2\tilde{\lambda},N)}{2}\right] \tag{143}$$

with

$$\widehat{\Psi}_{\pi,1}(2\tilde{\lambda},N) = \Psi_{\pi,1}(\tilde{\lambda},N) = \ln E_{f\sim\pi}\left(1 + \frac{2}{N}\frac{(m+1)!\left(6\tilde{\lambda}n_y\mu_{\max}(Q_e)G_e(f)^2\right)^2}{(1-6(m+1)\tilde{\lambda}n_y\mu_{\max}(Q_e)G_e(f)^2)}\right) \tag{144}$$

$$\widehat{\Psi}_{\pi,2}(2\tilde{\lambda},N) = \Psi_{\pi,2}(\tilde{\lambda},N) = \ln E_{f\sim\pi}\left(1 + \frac{(n_y+n_u)!}{\sqrt{N}}\frac{8\tilde{\lambda}\bar{G}_{gen}\bar{G}_f(f)}{1-8\tilde{\lambda}(n_y+n_u)\bar{G}_{gen}\bar{G}_f(f)}\right) \tag{145}$$

Proof A.17. we have

$$P(\omega \in S_1) \geq 1-\delta \tag{146}$$
$$P(\omega \in S_2) \geq 1-\delta \tag{147}$$

with

$$S_1 \triangleq \{\omega \in \Omega | \forall \rho \in \mathcal{M}_\pi : \quad E_{f\sim\hat{\rho}}\mathcal{L}(f) \leq E_{f\sim\hat{\rho}}V_N(f) + \frac{1}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \widehat{\Psi}_{\pi,1}(\lambda,N)\right]\} \tag{148}$$

$$S_2 \triangleq \{\omega \in \Omega | \forall \rho \in \mathcal{M}_\pi : \quad E_{f\sim\hat{\rho}}V_N(f) \leq E_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + \frac{1}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \widehat{\Psi}_{\pi,2}(\lambda,N)\right]\} \tag{149}$$

with $\bar{A}$ denoting the complementary set of A, i.e. $\bar{A} = \Omega \setminus A$

$$P(\omega \in \bar{S}_1) < \delta \tag{150}$$
$$P(\omega \in \bar{S}_2) < \delta \tag{151}$$
$$\tag{152}$$

Thus by union bound we get

$$P\left(\omega \in (\bar{S}_1 \cup \bar{S}_2)\right) < 2\delta \tag{153}$$

and thus

$$P\left(\omega \in (S_1 \cap S_2)\right) \geq 1-2\delta \tag{154}$$

with this we can write: with probability at least $1-2\delta$, the following holds

$$\forall \rho \in \mathcal{M}_\pi : \quad E_{f\sim\hat{\rho}}\mathcal{L}(f) \leq E_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + \frac{2}{\lambda}\left[KL(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \frac{\widehat{\Psi}_{\pi,2}(\lambda,N) + \widehat{\Psi}_{\pi,1}(\lambda,N)}{2}\right] \tag{155}$$

In order to bring this to a more common way of writing PAC-Bayesian bounds, let us define $\tilde{\lambda} = 0.5\lambda \leftrightarrow \lambda = 2\tilde{\lambda}$, thus we can write, for

$$0 < \tilde{\lambda} \leq \frac{1}{2}\Big(\sup_{f\in\mathcal{F}} \max\{3(m+1)n_y\mu_{\max}(Q_e)G_e(f)^2, 4n_w\bar{G}_{gen}\bar{G}_f(f)\}\Big)^{-1} \tag{156}$$

with probability at least $1-2\delta$, the following holds

$$\forall \rho \in \mathcal{M}_\pi : \quad E_{f\sim\hat{\rho}}\mathcal{L}(f) \leq E_{f\sim\hat{\rho}}\hat{\mathcal{L}}_N(f) + \frac{1}{\tilde{\lambda}}\left[KL(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \frac{\widehat{\Psi}_{\pi,2}(2\tilde{\lambda},N) + \widehat{\Psi}_{\pi,1}(2\tilde{\lambda},N)}{2}\right] \tag{157}$$

with

$$\widehat{\Psi}_{\pi,1}(2\tilde{\lambda},N)=\Psi_{\pi,1}(\tilde{\lambda},N)=\ln E_{f\sim\pi}\left(1+\frac{2}{N}\frac{(m+1)!\left(6\tilde{\lambda}n_y\mu_{\max}(Q_e)G_e(f)^2\right)^2}{(1-6(m+1)\tilde{\lambda}n_y\mu_{\max}(Q_e)G_e(f)^2)}\right) \tag{158}$$

$$\widehat{\Psi}_{\pi,2}(2\tilde{\lambda},N)=\Psi_{\pi,2}(\tilde{\lambda},N)=\ln E_{f\sim\pi}\left(1+\frac{(n_y+n_u)!}{\sqrt{N}}\frac{8\tilde{\lambda}\bar{G}_{gen}\bar{G}_f(f)}{1-8\tilde{\lambda}(n_y+n_u)\bar{G}_{gen}\bar{G}_f(f)}\right) \tag{159}$$

A.1. Bounded noise

In this section we state the lemmas and proofs associated with bounded innovation noise case.

Lemma A.22. *Let $\mathbf{e}_g(t)\in\mathcal{E}\subset\mathbb{R}^{n_y+n_y}$, be a zero mean, independant, and bounded stochastic process, s.t. $|\mathbf{e}_{g,i}(t)|\leq c_e$, $\forall i\in\{1,\dots,nu+ny\}$, i.e $\mathbf{e}_{g,i}(t)$ is the i'th component of $\mathbf{e}_g(t)$*

$$\mathbf{E}[\|\mathbf{e}_g(t)\|^r]\leq\left(c_e\sqrt{n_y+n_u}\right)^r \tag{160}$$

Proof A.18.

$$\mathbf{E}[\|\mathbf{e}_g(t)\|^r]=\mathbf{E}\left[\left(\sqrt{\sum_{i=1}^{nu+ny}\mathbf{e}_{g,i}^2(t)}\right)^r\right]\leq\left(\sqrt{\sum_{i=1}^{nu+ny}c_e^2}\right)^r=\left(\sqrt{(n_u+n_y)c_e^2}\right)^r=\left(c_e\sqrt{n_y+n_u}\right)^r \tag{161}$$

Lemma A.23. *Let $\mathbf{e}_g(t)\in\mathcal{E}\subset\mathbb{R}^{n_y+n_y}$, be a zero mean, independant, and bounded stochastic process, s.t. $|\mathbf{e}_{g,i}(t)|\leq c_e$, $\forall i\in\{1,\dots,nu+ny\}$, i.e $\mathbf{e}_{g,i}(t)$ is the i'th component of $\mathbf{e}_g(t)$*

$$\sigma(r)=\left(2c_e^2(n_y+n_u)\right)^r\geq\sup_{t,k,l}\mathbf{E}[\|e(t,k,l)\|_2^r] \tag{162}$$

$$e(t,k,l)=\mathbf{E}[\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)]-\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l) \tag{163}$$

Proof A.19. First let us take the case when $k\neq j$. Then, due to independance of $\mathbf{e}_g(t)$, we have $\mathbf{E}[\mathbf{e}_g(t-k)\mathbf{e}_g(t-j)]=0$, and thus

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]=\mathbf{E}[\|-\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j)\|_2^r]$$

Again as $\mathbf{e}_g(t)$ is i.i.d. we have

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]\leq\mathbf{E}[\left(\|\mathbf{e}_g(t-k)\|_2\|\mathbf{e}_g^T(t-j)\|_2\right)^r]\leq\mathbf{E}[\|\mathbf{e}_g(t-k)\|_2^r]\mathbf{E}[\|\mathbf{e}_g(t-j)\|_2^r]$$

and due to stationarity of $\mathbf{e}_g(t)$, we have $\mathbf{E}[\|\mathbf{e}_g(t-k)\|_2^r]=\mathbf{E}[\|\mathbf{e}_g(t-j)\|_2^r]$, therefore

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]\leq\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]^2$$

and again due to stationarity of $\mathbf{e}_g(t)$, the moments do not depend on t, and using Lemma A.22 we obtain

$$\forall k\neq j,\ \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]\leq\left(c_e^2(n_y+n_u)\right)^r$$

Now let us take the case when $k=j$. Then

$$\mathbf{E}\left[\|\mathbf{E}[\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)]-\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)\|^r\right]\leq\mathbf{E}\left[\left(\|\mathbf{E}[\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)]\|+\|\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)\|\right)^r\right] \tag{164}$$

By convexity $(a+b)^r=2^r\frac{1}{2^r}(a+b)^r=2^r\left(\frac{1}{2}(a+b)\right)^r\leq 2^{r-1}(a^r+b^r)$, we obtain

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]\leq 2^{r-1}\left(\mathbf{E}\left[\|\mathbf{E}[\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)]\|^r\right]+\mathbf{E}\left[\|\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)\|^r\right]\right) \tag{165}$$

$$=2^{r-1}\left(\|\mathbf{E}[\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)]\|^r+\mathbf{E}\left[\|\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)\|^r\right]\right) \tag{166}$$

$$\leq 2^{r-1}\left(\mathbf{E}[\|\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)\|^r]+\mathbf{E}\left[\|\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-l)\|^r\right]\right)\leq 2^r\mathbf{E}\left[\|\mathbf{e}_g(t)\|^{2r}\right] \tag{167}$$

Again by using Lemma A.22, we obtain

$$\forall k = j \ \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \left(2c_e^2(n_y+n_u)\right)^r \tag{168}$$

Thus we obtain the statement of the lemma

$$\forall t,k,j \ \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \max\{\left(c_e^2(n_y+n_u)\right)^r, \left(2c_e^2(n_y+n_u)\right)^r\} = \left(2c_e^2(n_y+n_u)\right)^r \tag{169}$$

Lemma A.24. *With notation as above, with $|\mathbf{e}_{g,i}| \leq c_e$, the following holds*

$$\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}] \leq 1 + \frac{1}{N} e^{\lambda 4 c_e^2 n_y (n_y+n_u) G_e(f)^2} \tag{170}$$

Proof A.20. By power series, and $\mathbf{E}[\mathcal{L}(f) - V_N(f)] = 0$, we have

$$\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}] = 1 + \sum_{r=2}^{\infty} \frac{\lambda^r}{r!} \mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \tag{171}$$

Now by Lemma A.15, and Lemma A.23, and $4(r-1) \leq 2^r$ we have

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \leq \frac{1}{N}(4c_e^2 n_y(n_y+n_u)G_e(f)^2)^r \tag{172}$$

Thus,

$$\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}] \leq 1 + \frac{1}{N}\sum_{r=2}^{\infty} \frac{1}{r!}(\lambda 4 c_e^2 n_y(n_y+n_u)G_e(f)^2)^r \tag{173}$$

now since $\lambda 4 c_e^2 n_y(n_y+n_u)G_e(f)^2 \geq 0$, then

$$1 + \frac{1}{N}\sum_{r=2}^{\infty} \frac{1}{r!}(\lambda 4 c_e^2 n_y(n_y+n_u)G_e(f)^2)^r \tag{174}$$

$$\leq 1 + \frac{1}{N}\sum_{r=0}^{\infty} \frac{1}{r!}(\lambda 4 c_e^2 n_y(n_y+n_u)G_e(f)^2)^r \tag{175}$$

$$= 1 + \frac{1}{N} e^{\lambda 4 c_e^2 n_y (n_y+n_u) G_e(f)^2} \tag{176}$$

Lemma A.25. *With notation as above, with $|\mathbf{e}_{g,i}| \leq c_e$, the following holds*

$$\mathbf{E}[e^{\lambda(V_N(f)-\hat{\mathcal{L}}(f))}] \leq 1 + \frac{1}{\sqrt{N}} e^{2\lambda G_f(f)\|\Sigma_{gen}\|_{\ell_1}^2 c_e^2(n_y+n_u)} \tag{177}$$

with

$$G_f(f) \triangleq \left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{(1-\hat{\gamma})^{\frac{3}{2}}}\right) \tag{178}$$

Proof A.21. By power series, we have

$$\mathbf{E}[e^{\lambda(V_N(f)-\hat{\mathcal{L}}(f))}] \leq \mathbf{E}[e^{\lambda|V_N(f)-\hat{\mathcal{L}}(f)|}] = 1 + \sum_{r=1}^{\infty} \frac{\lambda^r}{r!} \mathbf{E}[|V_N(f) - \hat{\mathcal{L}}(f)|^r] \tag{179}$$

For the terms $\mathbf{E}[|V_N(f) - \hat{\mathcal{L}}_N(f)|^r]$, we reuse the proof of Lemma A.12, and continue from (93), i.e.

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{2^r}{\sqrt{N}}\left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \sqrt{\frac{1}{1-\hat{\gamma}^{2r}}} \tag{180}$$

Note that

$$(1-\hat{\gamma})^{\frac{r}{2}} \leq \left(1-\hat{\gamma}^{2r}\right)^{\frac{1}{2}} \tag{181}$$

it is easy to see since for $\hat{\gamma} \in [0,1)$, the following holds

$$(1-\hat{\gamma})^{r} \leq 1-\hat{\gamma}^{2r} = (1-\hat{\gamma}^{r})(1+\hat{\gamma}^{r}) \tag{182}$$

$$1 \leq 1+\hat{\gamma}^{r} \tag{183}$$

This allows us to simplify the expression to

$$\mathbf{E}[\|V_N(f)-\hat{\mathcal{L}}_N(f)\|^r] \leq \frac{2^r}{\sqrt{N}}\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^r\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right]\left(\frac{1}{\sqrt{1-\hat{\gamma}}}\right)^r \tag{184}$$

Now, from Lemma A.6, we get

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \leq \|\Sigma_{gen}\|_{\ell_1}^{2r}\mathbf{E}[\|\mathbf{e}_g(t)\|^{2r}] \tag{185}$$

by lemma A.22, we get

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \leq \left(\|\Sigma_{gen}\|_{\ell_1}^2 c_e^2(n_y+n_u)\right)^r \tag{186}$$

Thus, with $G_f(f) \triangleq \frac{1}{\sqrt{1-\hat{\gamma}}}\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)$

$$\mathbf{E}[\|V_N(f)-\hat{\mathcal{L}}_N(f)\|^r] \leq \frac{1}{\sqrt{N}}\left(2G_f(f)\|\Sigma_{gen}\|_{\ell_1}^2 c_e^2(n_y+n_u)\right)^r \tag{187}$$

Thus

$$\mathbf{E}[e^{\lambda|V_N(f)-\hat{\mathcal{L}}(f)|}] \leq 1+\frac{1}{\sqrt{N}}\sum_{r=1}^{\infty}\frac{1}{r!}\left(2\lambda G_f(f)\|\Sigma_{gen}\|_{\ell_1}^2 c_e^2(n_y+n_u)\right)^r \tag{188}$$

$$\leq 1+\frac{1}{\sqrt{N}}e^{2\lambda G_f(f)\|\Sigma_{gen}\|_{\ell_1}^2 c_e^2(n_y+n_u)} \tag{189}$$

and therefore the statement of the lemma holds.

Corollary A.26. *By lemma A.18, lemmas A.24,A.25, and by applying a union bound, we obtain, for $\lambda > 0$, $\delta \in [0,1)$, the set of absolutely continuous probability density functions $\mathcal{M}_\pi$ w.r.t. π, the following holds with probability at least $1-2\delta$*

$$\forall \rho \in \mathcal{M}_\pi: \quad E_{f\sim\rho}\mathcal{L}(f) \leq E_{f\sim\rho}\hat{\mathcal{L}}_N(f)+\frac{1}{\lambda}\left[D_{\mathrm{KL}}(\rho||\pi)+\ln\left(\frac{1}{\delta}\right)+\widehat{\Psi}_{c_e,\pi}(\lambda,N)\right] \tag{190}$$

with

$$\widehat{\Psi}_{c_e,\pi}(\lambda,N) \triangleq \frac{1}{2}\left(\widehat{\Psi}_{c_e,\pi,1}(\lambda,N)+\widehat{\Psi}_{c_e,\pi,2}(\lambda,N)\right) \tag{191}$$

$$\widehat{\Psi}_{c_e,\pi,1}(\lambda,N) \triangleq \ln E_{f\sim\pi}\left(1+\frac{1}{N}e^{\lambda 4c_e^2 n_y(n_y+n_u)G_e(f)^2}\right) \tag{192}$$

$$\widehat{\Psi}_{c_e,\pi,2}(\lambda,N) \triangleq \ln E_{f\sim\pi}\left(1+\frac{1}{\sqrt{N}}e^{2\lambda G_f(f)\|\Sigma_{gen}\|_{\ell_1}^2 c_e^2(n_y+n_u)}\right) \tag{193}$$

A.2. Bounded innovation noise case: Alternative formulation

Lemma A.27. *for a sequence of random variables $x_j \in \mathbb{R}$, and $j \in \{1, \dots, r\}$*

$$\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \leq \left(\prod_{j=1}^{r-1} \mathbf{E}\left[x_j^{(2^j)}\right]^{(2^{-j})}\right) \mathbf{E}\left[x_r^{(2^{r-1})}\right]^{2^{-(r-1)}} \tag{194}$$

Proof A.22 (of Lemma A.27). We first apply Cauchy-Schwarz inequality $\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \leq |\mathbf{E}\left[\prod_{j=1}^{r} x_j\right]| = |\mathbf{E}\left[(x_1)\left(\prod_{j=2}^{r} x_j\right)\right]| \leq \sqrt{\mathbf{E}\left[x_1^2\right]}\sqrt{\mathbf{E}\left[\prod_{j=2}^{r} x_j^2\right]}$, and obtain

$$\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \leq \mathbf{E}\left[x_j^2\right]^{2^{-1}} \mathbf{E}\left[\prod_{j=2}^{r} x_j^2\right]^{2^{-1}} \tag{195}$$

Then we apply Cauchy-Schwarz again

$$\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \leq \mathbf{E}\left[x_1^2\right]^{2^{-1}} \mathbf{E}\left[x_2^{(2^2)}\right]^{2^{-2}} \mathbf{E}\left[\prod_{j=3}^{r} x_j^{(2^2)}\right]^{2^{-2}} = \prod_{j=1}^{2} \mathbf{E}\left[x_j^{(2^j)}\right]^{(2^{-j})} \mathbf{E}\left[\prod_{j=2+1}^{r} x_j^{(2^2)}\right]^{2^{-2}} \tag{196}$$

We repeat this process until we have

$$\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \leq \prod_{j=1}^{r-2} \mathbf{E}\left[x_j^{(2^j)}\right]^{(2^{-j})} \mathbf{E}\left[x_{r-1}^{(2^{r-2})} x_r^{(2^{r-2})}\right]^{2^{-(r-2)}} \tag{197}$$

Then we apply the final Cauchy-Schwarz inequality and obtain the statement of the lemma

$$\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \leq \prod_{j=1}^{r-2} \mathbf{E}\left[x_j^{(2^j)}\right]^{(2^{-j})} \mathbf{E}\left[x_{r-1}^{(2^{r-1})}\right]^{2^{-(r-1)}} \mathbf{E}\left[x_r^{(2^{r-1})}\right]^{2^{-(r-1)}} \tag{198}$$

$$= \prod_{j=1}^{r-1} \mathbf{E}\left[x_j^{(2^j)}\right]^{(2^{-j})} \mathbf{E}\left[x_r^{(2^{r-1})}\right]^{2^{-(r-1)}} \tag{199}$$

Lemma A.28. *Let $m = n_y + n_u$. If $|e_g(t)| < c_e$, then*

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\left(\frac{2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})}{N}\bar{G}_{f,2}(f)\right)^r \tag{200}$$

where $\bar{G}_{f,1}(f) \triangleq \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)$, and $\bar{G}_{f,2}(f) \triangleq \left(1 + \|\hat{D}\| + \frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\frac{1}{1-\hat{\gamma}}$ $\|\Sigma_{gen}\|_{\ell_1} \triangleq \|I\| + \sum_{k=1}^{\infty}\|C_g A_g^{k-1} K_g\|$.

Proof A.23 (of Lemma A.28). with $\mathbf{z}_\infty(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_f(t)$, and $\mathbf{z}_f(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_f(t|0)$, we start by applying triangle inequalities

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] = \mathbf{E}\left[\left|\frac{1}{N}\sum_{t=0}^{N-1}\|\mathbf{z}_\infty(t)\|^2 - \|\mathbf{z}_f(t)\|^2\right|^r\right] \leq \mathbf{E}\left[\left(\frac{1}{N}\sum_{t=0}^{N-1}\left|\|\mathbf{z}_\infty(t)\|^2 - \|\mathbf{z}_f(t)\|^2\right|\right)^r\right] \tag{201}$$

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\mathbf{E}\left[\prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t_j)\|^2 - \|\mathbf{z}_f(t_j)\|^2\right|\right] \tag{202}$$

Now using the fact that $|a^2 - b^2| = |(a-b)(a+b)| = |a-b|(a+b)$, since $a, b \geq 0$, we get

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_r=0}^{N-1} \mathbf{E}\left[\prod_{j=1}^{r} |\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\|| \left(\|\mathbf{z}_\infty(t_j)\| + \|\mathbf{z}_f(t_j)\|\right)\right] \tag{203}$$

We apply Cauchy-Schwarz, i.e. $\mathbf{E}[XY] \leq |\mathbf{E}[XY]| \leq \sqrt{\mathbf{E}[X^2]}\sqrt{\mathbf{E}[Y^2]}$, with $X = \prod_{j=1}^{r} |\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\||$, and $Y = \prod_{j=1}^{r} \left(\|\mathbf{z}_\infty(t_j)\| + \|\mathbf{z}_f(t_j)\|\right)$,

$$\mathbf{E}[\|V_N(f) - \hat{\mathcal{L}}_N(f)\|^r] \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_r=0}^{N-1} \sqrt{\mathbf{E}\left[\prod_{j=1}^{r} |\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\||^2\right]} \sqrt{\mathbf{E}\left[\prod_{j=1}^{r} \left(\|\mathbf{z}_\infty(t_j)\| + \|\mathbf{z}_f(t_j)\|\right)^2\right]} \tag{204}$$

For now let's focus on $\mathbf{E}\left[\prod_{j=1}^{r} |\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\||^2\right]$, by applying reverse triangle inequality we obtain

$$\mathbf{E}\left[\prod_{j=1}^{r} |\|\mathbf{z}_\infty(t_j)\| - \|\mathbf{z}_f(t_j)\||^2\right] \leq \mathbf{E}\left[\prod_{j=1}^{r} \|\mathbf{z}_\infty(t_j) - \mathbf{z}_f(t_j)\|^2\right] \tag{205}$$

For the ease of notation for the next step, let us define $x_j \triangleq \|\mathbf{z}_\infty(t_j) - \mathbf{z}_f(t_j)\|^2$, then the quantity of interest is

$$\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \tag{206}$$

For the above quantity we can apply Lemma A.27, which states

$$\mathbf{E}\left[\prod_{j=1}^{r} x_j\right] \leq \prod_{j=1}^{r-1} \mathbf{E}\left[x_j^{(2^j)}\right]^{(2^{-j})} \mathbf{E}\left[x_r^{(2^{r-1})}\right]^{2^{-(r-1)}} \tag{207}$$

From Lemma A.7, we also know that

$$\mathbf{E}[\|\mathbf{z}_\infty(t) - \mathbf{z}_f(t)\|^r] \leq \hat{\gamma}^{rt} \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \tag{208}$$

Thus combining Lemma A.27 and Lemma A.7, we get

$$\mathbf{E}\left[\prod_{j=1}^{r} \|\mathbf{z}_\infty(t_j) - \mathbf{z}_f(t_j)\|^2\right] \leq \prod_{j=1}^{r-1} \hat{\gamma}^{2t_j} \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^2 \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{(2^{j+1})}\right]^{\frac{1}{2^j}}$$
$$\times \hat{\gamma}^{2t_r} \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^2 \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^{(2^r)}\right]^{\frac{1}{2^{r-1}}} \tag{209}$$

with Lemma A.10, and Lemma A.22, we have

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|^r\right] \leq \|\Sigma_{gen}\|_{\ell_1}^r (c_e\sqrt{m})^r, \tag{210}$$

thus we get

$$\mathbf{E}\left[\prod_{j=1}^{r} \|\mathbf{z}_\infty(t_j) - \mathbf{z}_f(t_j)\|^2\right] \leq \prod_{j=1}^{r-1} \hat{\gamma}^{2t_j} \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^2 \left(\|\Sigma_{gen}\|_{\ell_1}^{2^{j+1}} (c_e\sqrt{m})^{2^{j+1}}\right)^{2^{-j}}$$
$$\cdot \hat{\gamma}^{2t_r} \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^2 \left(\|\Sigma_{gen}\|_{\ell_1}^{2r} (c_e\sqrt{m})^{2r}\right)^{2^{-(r-1)}}, \tag{211}$$

With some algebraic simplification we obtain the first term

$$\mathbf{E}\left[\prod_{j=1}^{r}\|\mathbf{z}_\infty(t_j)-\mathbf{z}_f(t_j)\|^2\right]\leq\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^2\|\Sigma_{gen}\|_{\ell_1}^2(c_e\sqrt{m})^2\prod_{j=1}^{r}\hat{\gamma}^{2t_j}, \tag{212}$$

Now for the second term $\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^2\right]$, we apply the inequality of arithmetic-geometric means

$$\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^2\right]\leq\frac{1}{r}\sum_{j=1}^{r}\mathbf{E}\left[\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^{2r}\right] \tag{213}$$

By Lemma A.11, we obtain

$$\frac{1}{r}\sum_{j=1}^{r}\mathbf{E}\left[\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^{2r}\right]\leq\frac{2^{2r-1}}{r}\sum_{j=1}^{r}\left(\mathbf{E}\left[\|\mathbf{z}_\infty(t_j)\|^{2r}\right]+\mathbf{E}\left[\|\mathbf{z}_f(t_j)\|^{2r}\right]\right) \tag{214}$$

By Lemma A.8 and Lemma A.9, we obtain

$$\frac{2^{2r-1}}{r}\sum_{j=1}^{r}\left(\mathbf{E}\left[\|\mathbf{z}_\infty(t_j)\|^{2r}\right]+\mathbf{E}\left[\|\mathbf{z}_f(t_j)\|^{2r}\right]\right)\leq\frac{2^{2r}}{r}\sum_{j=1}^{r}\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^{2r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \tag{215}$$

$$=2^{2r}\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)^{2r}\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{2r}\right] \tag{216}$$

with Lemma A.10, and Lemma A.22, we have

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|^{r}\right]\leq\|\Sigma_{gen}\|_{\ell_1}^{r}(c_e\sqrt{m})^r, \tag{217}$$

we get

$$\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^2\right]\leq\left(2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\right)^{2r} \tag{218}$$

Now taking (218) and (80) back to (204), we have

$$\begin{aligned}\mathbf{E}[\|V_N(f)-\hat{\mathcal{L}}_N(f)\|^r]&\leq\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\mathbf{E}\left[\prod_{j=1}^{r}\left|\|\mathbf{z}_\infty(t_j)\|-\|\mathbf{z}_f(t_j)\|\right|^2\right]}\sqrt{\mathbf{E}\left[\prod_{j=1}^{r}\left(\|\mathbf{z}_\infty(t_j)\|+\|\mathbf{z}_f(t_j)\|\right)^2\right]}\\&\leq\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sqrt{\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)^2\|\Sigma_{gen}\|_{\ell_1}^2(c_e\sqrt{m})^2\prod_{j=1}^{r}\hat{\gamma}^{2t_j}}\\&\quad\cdot\sqrt{\left(2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\right)^{2r}}\end{aligned} \tag{219}$$

with $G_f(f)\triangleq\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)$

$$\begin{aligned}\mathbf{E}[\|V_N(f)-\hat{\mathcal{L}}_N(f)\|^r]&\leq\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\left(2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\right)^r\\&\quad\cdot\frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\prod_{j=1}^{r}\hat{\gamma}^{t_j}\end{aligned} \tag{220}$$

Note that $\left(\sum_{t=0}^{N-1}\hat{\gamma}^t\right)^r = \sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\prod_{j=1}^{r}\hat{\gamma}^{t_j}$, and by applying the sum of the geometric series we obtain

$$\mathbf{E}[\|V_N(f)-\hat{\mathcal{L}}_N(f)\|^r] \leq \left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\left(2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\right)^r \cdot\left(\frac{1-\hat{\gamma}^N}{N(1-\hat{\gamma})}\right)^r \tag{221}$$

Note that $1-\hat{\gamma}^N\leq 1$, so with $\bar{G}_{f,1}(f)\triangleq\left(\frac{\hat{M}\|\hat{C}\|\|\hat{B}\|}{1-\hat{\gamma}}\right)$, and $\bar{G}_{f,2}(f)\triangleq\left(1+\|\hat{D}\|+\frac{\hat{M}\|\hat{B}\|\|\hat{C}\|}{1-\hat{\gamma}}\right)\frac{1}{1-\hat{\gamma}}$ the statement of the lemma follows.

Lemma A.29. *With notation as above the following holds*

$$\mathbf{E}[e^{\lambda|V_N(f)-\hat{\mathcal{L}}_N(f)|}] \leq 1+\bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\sum_{r=1}^{\infty}\frac{\left(\lambda\frac{2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})}{N}\bar{G}_{f,2}(f)\right)^r}{r!} \tag{222}$$

$$=(1-\bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m}))+\bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})e^{\lambda\frac{2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})}{N}\bar{G}_{f,2}(f)}$$

Proof A.24 (of Lemma A.13). with $X=\lambda|V_N(f)-\hat{\mathcal{L}}_N(f)|$

$$\mathbf{E}[e^{\lambda(V_N(f)-\hat{\mathcal{L}}_N(f))}] = 1+\sum_{r=1}^{\infty}\frac{\lambda^r}{r!}\mathbf{E}[|V_N(f)-\hat{\mathcal{L}}_N(f)|^r] \leq 1+\sum_{r=1}^{\infty}\frac{\lambda^r}{r!}\left(\frac{2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})}{N}\bar{G}_{f,2}(f)\right)^r \tag{223}$$

Lemma A.30 (Alternative bound using (Alquier & Wintenberger, 2012))**.** *With probability at least* $1-\delta$*, the following holds*

$$\forall\rho:\quad E_{f\sim\hat{\rho}}\mathcal{L}(f)\leq\ E_{f\sim\hat{\rho}}V_N(f)+\frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}\|\pi)+\ln\frac{1}{\delta}+\Psi_{\pi,2}(\lambda,N)\right], \tag{224}$$

with

$$\Psi_{\pi,2}(\lambda,N)=\ln E_{f\sim\pi}\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}]\leq\ln E_{f\sim\pi}\left(e^{\frac{\lambda^2}{2N}(G_e(f)+G_{e,1}(f))^2C^2(4G_e(f)C+1)^2}\right) \tag{225}$$

where $C=c_e\sqrt{n_u+n_y}$

$$G_{e,1}(f)=\|D_e\|_2+\sum_{k=1}^{\infty}(k+1)\|C_eA_e^{k-1}K_e\|_2$$

In particular, $\lim_{N\to\infty}\Psi_{\pi,2}(\lambda,N)=0$ *for any* $\lambda>0$ *and for* $\lambda_N=\sqrt{N}$, $\lim_{N\to}\frac{1}{\lambda_N}\Psi_{\pi,2}(\lambda_N,N)=0$.

Proof A.25 (Proof of Lemma A.30). For each $f\in\mathcal{F}$, consider $\mathbf{X}_t=\mathbf{y}(t)-\hat{\mathbf{y}}_f(t)$. Then $\mathbf{X}_t$

$$\mathbf{X}_t=\sum_{k=0}^{\infty}\alpha_k\mathbf{e}_g(t-k),$$

where

$$\alpha_k=\begin{cases}D_e, & k=0\\ C_eA_e^{k-1}K_e, & k>0\end{cases}$$

By (Alquier et al., 2013, Proposition 4.2) X_t is a weakly dependent process in the terminology of (Alquier et al., 2013), and $\|X_t\|\leq G_e(f)C$ and the coefficient $\theta_{\infty,N}(1)$ satisfies $\theta_{\infty,N}(1)<2G_{e,1}(f)C$ for all $N\mathbb{N}$. Consider the function $h(x_1,\ldots,x_N)=\frac{1}{(2L+1)}\sum_{i=1}^{N}\|x_i\|_2^2$ defined on $\mathcal{X}=[-L,L]^N$, where $L=2G_e(f)C$. Then h is $1-Lipschitz$. Notice that $\lambda V_N(f)=\frac{\lambda}{N}(2L+1)h(\mathbf{X}(0),\ldots,\mathbf{X}(N-1))$. Then

$$\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}]=\mathbf{E}[e^{\frac{\lambda}{N}(2L+1)(\mathbf{E}[h(\mathbf{X}(0),\ldots,\mathbf{X}(N-1))]-h(\mathbf{X}(0),\ldots,\mathbf{X}(N-1))}]$$

and hence by (Alquier et al., 2013, Theorem 6.6)

$$\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}] \leq e^{\frac{\lambda^2}{N}(2L+1)^2(\|\mathbf{X}_0\|_\infty+\theta_{\infty,N}(1))^2/2}$$

where $\|\mathbf{X}_0\|_\infty$ is the smallest real number such that $\|\mathbf{X}_0\| \leq \|\mathbf{X}_0\|_\infty$ with probability 1. By using the definition L, and the facts that $\|X_t\| \leq G_e(f)C$ and $\theta_{\infty,N}(1) < 2G_{e,1}(f)C$ the statement of the lemma follows.

$$\mathbf{E}[e^{\lambda(\mathcal{L}(f)-V_N(f))}] \leq e^{\frac{\lambda^2}{N}(2L+1)^2(\|\mathbf{X}_0\|_\infty+\theta_{\infty,N}(1))^2/2} \leq e^{\frac{\lambda^2}{N}(4G_e(f)C+1)^2(G_e(f)+2G_{e,1})^2C^2/2} \tag{226}$$

Proof A.26 (of Theorem 5.4). By applying Lemma A.30, Lemma A.29, and by applying the union bound as in Lemma A.21, we obtain, for $\lambda > 0$, $\delta \in (0, 1]$, with probability at least $1 - 2\delta$

$$\forall \rho \in \mathcal{M}_\pi : \; E_{f\sim\rho}\mathcal{L}(f) \leq \mathbb{E}_{f\sim\rho}\hat{\mathcal{L}}_N(f) + \frac{2}{\lambda}\left[D_{\mathrm{KL}}(\rho|\pi) + \ln\frac{1}{\delta} + \frac{\Psi_1(\lambda, N) + \Psi_2(\lambda, N)}{2}\right] \tag{227}$$

with

$$\Psi_1(\lambda, N) \triangleq \ln E_{f\sim\pi} e^{\frac{\lambda^2}{2N}(4G_e(f)C+1)^2(G_e(f)+2G_{e,1})^2C^2} \tag{228}$$

$$\Psi_2(\lambda, N) \triangleq \ln E_{f\sim\pi}\left((1 - \bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})) + \bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})e^{\lambda\frac{2\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})}{N}\bar{G}_{f,2}(f)}\right) \tag{229}$$

Now with $\tilde{\lambda} \triangleq 0.5\lambda \leftrightarrow \lambda = 2\tilde{\lambda}$, we obtain the statement of the lemma: for $\tilde{\lambda} > 0$, $\delta \in (0, 1]$, then with probability at least $1 - 2\delta$

$$\forall \rho \in \mathcal{M}_\pi : \; E_{f\sim\rho}\mathcal{L}(f) \leq \mathbb{E}_{f\sim\rho}\hat{\mathcal{L}}_N(f) + \frac{1}{\tilde{\lambda}}\left[D_{\mathrm{KL}}(\rho|\pi) + \ln\frac{1}{\delta} + \frac{\Psi_1(\tilde{\lambda}, N) + \Psi_2(\tilde{\lambda}, N)}{2}\right] \tag{230}$$

with

$$\Psi_1(\tilde{\lambda}, N) \triangleq \ln E_{f\sim\pi} e^{\frac{\tilde{\lambda}^2}{N}2(4G_e(f)C+1)^2(G_e(f)+2G_{e,1})^2C^2} \tag{231}$$

$$\Psi_2(\tilde{\lambda}, N) \triangleq \ln E_{f\sim\pi}\left((1 - \bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})) + \bar{G}_{f,1}(f)\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})e^{\frac{\tilde{\lambda}}{N}8\|\Sigma_{gen}\|_{\ell_1}(c_e\sqrt{m})\bar{G}_{f,2}(f)}\right) \tag{232}$$